Uniform Convergence and Learnability

by

Martin Henry George Anthony
London School of Economics

February 1991

A Dissertation for the Degree of
Doctor of Philosophy
of
The University of London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis analyses some of the more mathematical aspects of the Probably Approximately Correct (PAC) model of computational learning theory.

The main concern is with the sample size required for valid learning in the PAC model. A sufficient sample-size involving the Vapnik-Chervonenkis (VC) dimension of the hypothesis space is derived; this improves the best previously known bound of this nature.

Learnability results and sufficient sample-sizes can in many cases be derived from results of Vapnik on the uniform convergence (in probability) of relative frequencies of events to their probabilities, when the collection of events has finite VC dimension. Two simple new combinatorial proofs of each of two of Vapnik's results are proved here and the results are then applied to the theory of learning stochastic concepts, where again improved sample-size bounds are obtained.

The PAC model of learning is a distribution-free model; the resulting sample sizes are not permitted to depend on the usually fixed but unknown probability distribution on the input space. Results of Ben-David, Benedek and Mansour are described, presenting a theory for distribution-dependent learnability. The conditions under which a feasible upper bound on sample-size can be obtained are investigated, introducing the concept of polynomial $X\sigma$-finite dimension.

The theory thus far is then applied to the learnability of formal concepts, defined by Wille. A learning algorithm is also presented for this problem.

Extending the theory of learnability to the learnability of functions which have range in some arbitrary set, learnability results and sample-size bounds, depending on a generalization of the VC dimension, are obtained and these results are applied to the theory of artificial neural networks. Specifically, a sufficient sample-size for valid generalization in multiple-output feedforward linear threshold networks is found.

Contents

Acknowledgements

I have benefitted greatly from the kind assistance and encouragement of many people throughout my work on this thesis.

In particular, Prof. Norman Biggs, my supervisor, has been a source of great encouragement. I am indebted to him for his help and his advice, always generously given whenever requested.

I also thank Dr. John Shawe-Taylor for his encouragement and enthusiasm, and for the many fruitful and enjoyable discussions we have had.

The Mathematics departments of Royal Holloway and Bedford New College (University of London) and the London School of Economics provided friendly, comfortable and stimulating environments in which to work and contributed greatly to making my time as a research student so enjoyable. In particular, I thank the postgraduate mathematicians at RHBNC and Dr. Graham Brightwell at LSE.

This work would not have been possible had the Science and Engineering Research Council not provided financial support for the first two years. I thank the Council for this support, and for their generous financial assistance with a visit to the USA.

I am grateful to Prof. Lenny Pitt of the University of Illinois for his hospitality and for giving so freely of his time.

I also wish to thank Mick Ganley, John Livingstone, Audrey Marshall, Colleen McKenna and Mark Overend, a few of the friends and former teachers who have provided me with great encouragement and support.

Finally, I wish to thank my parents, Duncan Anthony and Jean Cameron Anthony, for their immense and selfless support over the years. I dedicate this work to them.

Declaration of Originality

I declare that, with the exceptions detailed below, this thesis is my own unaided work.

I have tried to make it clear in the text when a result is not new or not due to me (even if the proof is original). I have also attempted throughout to reference work that has motivated my research. Some areas of this thesis are the result of joint work to which I made a very significant contribution.

Chapter 1 is due to me and is largely introductory.

Chapter 2 contains some standard results. Dr. Graham Brightwell helped with the proof of Theorem 2.8. All other new results and new proofs, as specified in the Preface, are due to me.

Chapter 3 extends joint work with Prof. Norman Biggs and Dr. John Shawe-Taylor. The new ideas were developed jointly, and a weaker upper bound on sufficient sample-size was found. I have extended the analysis. The presentation and the sample-size bound here are due to me. Section 3.1 is introductory. Sections 3.2 and 3.3 are due to me, with the exception of Lemma 3.1, which is due to Dr. Graham Brightwell. These sections are based upon the joint work with Dr. Shawe-Taylor and Prof. Biggs. Section 3.4 is due to me, except (as stated in the text) for Theorems 3.18 and 3.19.

Chapters 4 and 5 are due to me.

Chapter 6 is, in part, joint work with Dr. John Shawe-Taylor. Sections 6.1 and 6.2 are introductory. In Section 6.3, the proof of Proposition 6.6 is joint work with Dr. Shawe-Taylor. The subsection on hypothesis spaces of $X\sigma$-finite dimension contains an account of work by Ben-David, Benedek and Mansour. The subsection on polynomial distribution-dependent learnability is joint work with Dr. Shawe-Taylor. The discussion and examples after Theorem 6.13 are due to me.

Sections 7.1 and 7.2, with the exception of the last subsection, supply an account of standard notions of Formal Concept Analysis. The subsection on Formal concepts and monomials is original and due to me. Initially I proved a weaker form of Theorem 7.9 of section 7.3, and this was circulated in a preprint written jointly with Prof. Norman Biggs. I am grateful to Dr. Colin McDiarmid for suggesting that the stronger statement holds with essentially the same proof. Theorem 7.11 is, as stated, a result of Dr. Dave Cohen.

Chapter 8 is due to me.

Chapter 9 is, in part, joint work with Dr. John Shawe-Taylor. Section 9.1 is introductory. Sections 9.2 and 9.3 are due to me. Section 9.4 is joint work with Dr. Shawe-Taylor.

Declaration of Originality

I declare that, with the exceptions detailed below, this thesis is my own unaided work.

I have tried to make it clear in the text when a result is not new or not due to me (even if the proof is original). I have also attempted throughout to reference work that has motivated my research. Some areas of this thesis are the result of joint work to which I made a very significant contribution:

The upper bound on sufficient sample-size in Chapter 3 improves upon joint research with Prof. Norman Biggs and Dr. John Shawe-Taylor, and the analysis extends that work.

Proposition 6.6 and the treatment of polynomial $X\sigma$-finite dimension is joint work with Dr. John Shawe-Taylor.

Section 9.4 is joint work with Dr. John Shawe-Taylor.

In addition, Lemma 3.1 is due to Dr. Graham Brightwell, and his suggestions facilitated the proof of Theorem 2.8.

Preface

This thesis studies the sample complexity of Valiant's Probably Approximately Correct model of learning, together with related problems in probability theory, and applies the results to the theory of formal concepts and to the theory of learning in artificial neural networks.

Chapters 1 and 2 are introductory. Chapter 1 introduces the idea of PAC learnability and sets up the essential definitions. Chapter 2 discusses the Vapnik-Chervonenkis dimension for collections of sets and Boolean-valued functions. It contains a new treatment of the VC dimension of half-spaces of Euclidean space, a new bound on the number of Boolean threshold functions on a given number of variables, and a result which extends a result of Haussler and Welzl concerning the VC dimension of a graph.

In Chapter 3, we obtain bounds on the sample-size required for learnability, these bounds depending on the Vapnik-Chervonenkis dimension of the hypothesis space. First, a result is obtained which bounds the probability of presenting a "bad" training sample. This involves the expected values of the index functions, which we show exist if the hypothesis space is universally separable. The result is applied to obtain an upper bound on sufficient sample-size which is better than previously obtained bounds. We end the chapter by discussing lower bounds on necessary sample-size.

Chapter 4 concerns the uniform convergence (in probability) of the relative frequencies of events in some class of finite VC dimension to their probabilities. This is an area of probability theory which underpins learnability theory (and in particular sample-sizes). Two simple new combinatorial proofs of each of

two results of Vapnik are given. We also give a quick and easy proof of a learnability result of Haussler, derived in the previous chapter as a corollary of the more general analysis presented there.

In Chapter 5, we apply the uniform convergence results to obtain results on the approximation of stochastic concepts by spaces of Boolean-valued functions of finite VC dimension. This idea is not new, but we discuss the measurability aspects and the resulting sample-size improves upon the best previously obtained. We end the chapter with a brief discussion of possible applications of this theory.

We discuss non-uniform learnability in Chapter 6, describing sufficient conditions on a hypothesis space for it to be learnable in a distribution-dependent manner. An important problem is to guarantee not merely a finite but a feasibly small sufficient sample-size. The concept of polynomial $X\sigma$-finite dimension with respect to a particular distribution is introduced and is shown to imply distribution-dependent learnability with feasible sample-size.

In Chapter 7, we apply the theory of learnability to Wille's formal concept analysis, showing that in contexts having certain boundedness properties, the set of formal concept extents has bounded VC dimension and is thus learnable. An algorithm for learning formal concept extents in a finite context is presented.

Chapter 8 discusses the generalizations of VC dimension and learnability to spaces of functions which have general range. We describe how a generalized VC dimension may be defined for such spaces and we define stochastic concepts with range in some countable set. The results of previous chapters are then applied to give learnability results for such functions and such stochastic concepts.

Chapter 9 uses the framework and results of Chapter 8 to study the sample-size required for valid generalization in certain types of artificial neural network. We briefly define and discuss artificial neural networks and discuss previous

results on sample-size. A bound is obtained on the (generalized) VC dimension of the space of functions computable by a feedforward linear threshold network with real inputs. This result leads to a sample-size upper bound which depends on the number of computation nodes and the number of weights but not on the number of output nodes. This extends a result of Baum and Haussler on feedforward linear threshold networks with a single output node.

Chapter 1
PAC Learning and Learnability

1.1 Introduction

In this introductory chapter we describe Valiant's Probably Approximately Correct (PAC) model of learning. This is a stochastic model in which a hypothesis, from a set of hypotheses, is chosen which is meant to approximate to a target concept, usually also from the same set of hypotheses. It is shown that PAC learning can be achieved if there is an efficient *consistent hypothesis finder* and if the hypothesis space is *potentially learnable.*

We then formalise potential learnability further, describing the measurability constraints to be imposed upon the spaces we consider. We end the chapter with a proof that any finite hypothesis space is potentially learnable.

1.2 Probably Approximately Correct Learning

Motivation and informal definitions

A few years ago Valiant [33, 34] described a computational model of learning which Angluin [1] has called the Probably Approximately Correct (or PAC) learning model.

Suppose that a set X of objects is partitioned into two sets, called the positive and the negative examples. We think of this partition as representing a concept or a classification of the objects; formally, the concept is the set of positive examples or the characteristic function of this set. In what follows, we often identify subsets with $\{0,1\}$-valued functions; given a subset, we may represent

it by its characteristic function, and given a boolean-valued function, we may identify it with its support. A learner, which may be thought of as a machine or algorithm, has to choose a $\{0,1\}$-function (from a given set of functions) which is supposed to approximate to the concept. The only information given to the learner is the set of functions available to choose from, a finite sequence of objects and information as to whether each of the objects in this sequence is a positive example or a negative example of the concept being learned. The PAC model is a stochastic model of learning in which it is required that, in a probabilistic sense to be made precise below, the learner generalizes well from the examples presented to it during the training procedure. The description we give below is framed in terms of boolean-valued functions, but can be modified in the obvious way to refer to subsets of the set of inputs.

Roughly speaking, the idea of PAC learning is that a function

$$c : X \to \{0,1\}$$

(the *target function* or *target concept*) from H is *PAC learnable* by a set of functions H (the *hypothesis space*) if there is an algorithm $\mathcal{L}$ (the *learning algorithm*) which takes as input a sequence of randomly chosen elements of X, labelled with the values of c on these elements (a *training sample*), and returns (a representation of) a hypothesis $h \in H$ such that the following holds:

For a sufficiently large training sample, there is high probability that the resulting hypothesis is approximately equal to c.

The *input space* X is assumed to have a fixed probability measure μ defined on it and, for a sample of length m, the probability referred to above is the product probability measure μ^m on X^m. This is the distribution from which the training sample is randomly drawn. Since the measure is generally unknown, we require that the above condition holds for any probability measure μ on X. Additionally, since the target concept c is not known to the learner, we require that the above condition holds for any c in H. The "sufficiently large" sample-size above must not depend an the distribution or on the target

concept, for neither of these is known by the learner: In attempting to learn some target concept from H, the learner has to be able to guarantee that if it takes a certain number of randomly drawn training examples, it will probably output a good approximation to the target concept. This guarantee clearly cannot be made in general if the sufficient length of training sample depends on things unknown to the learner.

Formal definition

We define the *actual error* $\mathrm{er}_\mu(h, c)$ of the hypothesis h with respect to c and μ to be the probability that on a further randomly chosen input, h and c disagree. That is,

$$\mathrm{er}_\mu(h, c) = \mu\{x \in X : h(x) \neq c(x)\}.$$

(Assume for the moment that this measure is defined; we address measurability conditions later in the chapter).

We can formally define a training sample from X of a hypothesis from H. Suppose that $c \in H$ and that

$$(x_1, x_2, \ldots, x_m) \in X^m.$$

Then the *training sample* $c(\mathrm{x})$ of c on x is the vector x labelled with the values of $c(x_1), \ldots, c(x_m)$. That is,

$$c(\mathrm{x}) = ((x_1, c(x_1)), (x_2, c(x_2)), \ldots, (x_m, c(x_m))).$$

Then the above informal description of PAC learning a particular concept may be formalised and used to define the PAC learnability of the hypothesis space H. (Recall that, by the above discussion, we require every hypothesis from H to be learnable and the sufficient sample-size to be independent of both the distribution and the particular hypothesis chosen to be the target concept).

Definition 1.1 *The hypothesis space H is PAC learnable if there is an algorithm $\mathcal{L}$ taking as input training samples from X of hypotheses from H such that the following holds: Given an accuracy parameter $0 < \epsilon < 1$ and a confidence parameter $0 < \delta < 1$, there is a positive integer $m_0 = m_0(\epsilon, \delta)$ (a sufficient sample-size) such that*

$$m \geq m_0 \Longrightarrow \mu^m \{x \in X^m : \mathrm{er}_\mu(\mathcal{L}(c(x)), c) < \epsilon\} > 1 - \delta,$$

for any probability measure μ on X and for any c in H. □

Complexity of the learning algorithm

Usually, some complexity conditions must be imposed on the learning algorithm $\mathcal{L}$. For feasible computation, $\mathcal{L}$ should run in a time which is polynomial in various parameters characterizing the complexity of the particular learning problem. In particular, $\mathcal{L}$ should run in a time polynomial in $1/\epsilon$ and $1/\delta$, where ϵ and δ are (respectively) the accuracy and confidence parameters. This complexity condition can be met if $\mathcal{L}$ runs in a time polynomial in its input and if the sufficient sample-size $m_0(\epsilon, \delta)$ is polynomial in $1/\epsilon$ and $1/\delta$.

If $\mathcal{L}$ is a learning algorithm for H which runs in a time polynomial in $1/\epsilon$ and $1/\delta$, we say H is *polynomially learnable* by $\mathcal{L}$. (More generally, it is often appropriate that the algorithm should be required to operate in a time polynomial in some measure of the complexity or size of both the input space and the target concept. See, for example [26]).

1.3 Potential Learnability

Potential learnability and consistent hypothesis finders

Informally, notice that H will certainly be PAC learnable if for any $c \in H$:

• With probability close to 1, any hypothesis from H consistent with c on sufficiently many randomly chosen inputs from X is approximately equal to c; and

• There is an algorithm $\mathcal{L}$ for finding a hypothesis from H consistent with c on any training sample from X of a hypothesis from H. Such an algorithm is called a *consistent hypothesis finder.*

The first condition leads to the following definition of *potential learnability.*

Definition 1.2 *The hypothesis space H is said to be potentially learnable if given $\epsilon > 0$ and $\delta > 0$ there is an integer $m_0 = m_0(\epsilon, \delta)$ such that for all $m \geq m_0$,*

$$\mu^m \{(x_1, x_2, \ldots, x_m) \in X^m : \forall h \in H,\, t(x_i) = h(x_i) \forall i \Rightarrow \mathrm{er}_\mu(h) < \epsilon\} > 1 - \delta,$$

for any probability measure μ defined on X, and for all $t \in H$. □

We have not stipulated in this definition that $m_0(\epsilon, \delta)$ be polynomial in $1/\epsilon$ and $1/\delta$ (a desirable property, by the discussion in the previous section). In fact, as a consequence of results presented later, if H is potentially learnable then one can find a value of $m_0(\epsilon, \delta)$ which is polynomial in $1/\epsilon$ and $1/\delta$.

It is not the aim here to discuss the algorithmic aspects of PAC learning, an area which has interested many researchers in recent years, and in which much work remains to be done (see, for example, [26]). Rather, we shall be concerned with the size of sample required for learning; that is, the sample complexity or information complexity of learning. For this reason, we study potential learnability and will refer to potential learnability from now on simply as *learnability.*

1.4 The Input Space and the Hypothesis Space

We now specify the types of input space and hypothesis space we consider. These are obtained by imposing certain measure-theoretic and non-triviality conditions.

The input space as a probability space

Throughout, the input space X is assumed to be either finite, countably infinite or a subset of Euclidean space $\mathbf{R}^n$ for some n. In the first two of these cases, we take the σ-algebra Σ to be 2^X, the set of all subsets of X. If X is a subset of the Euclidean space $\mathbf{R}^n$, we take Σ to be the Borel σ-algebra induced on X; that is

$$\Sigma = \{B \cap X : B \in \mathcal{B}\},$$

where $\mathcal{B}$ is the σ-algebra of subsets of $\mathbf{R}^n$ generated by the subsets open with respect to (for example) the standard Euclidean metric. (An alternative description of this latter σ-algebra is as the σ-algebra of subsets of X generated by the subsets of X which are open with respect to the induced Euclidean metric on X.)

The probability distribution according to which examples are drawn is a probability measure μ defined on the σ-algebra Σ. When we discuss probability measures μ on X, we shall mean probability measures μ on the σ-algebra Σ (where Σ is defined as above for the various cases); that is, measures such that (X, Σ, μ) is a probability space. In particular, a statement of the form "For all probability measures μ on X..." should be interpreted as "For all probability measures μ defined on the σ-algebra Σ, where Σ is the power set of X if X is countable and is the induced Borel σ-algebra if X is Euclidean, ...".

The hypothesis space

We assume that H is a set of Σ-measurable functions from X to $\{0,1\}$. This is equivalent to demanding that for each h in H, the set $h^{-1}(1)$ belongs to Σ. With this measurability condition, it is possible as earlier to define the *(actual) error* of one hypothesis with respect to another hypothesis as the measure of the set of inputs on which they disagree, this set being measurable. That is, the error $\mathrm{er}_\mu(h,c)$ of $h \in H$ with respect to $c \in H$ is

$$\mathrm{er}_\mu(h,c) = \mu\{x \in X : h(x) \neq c(x)\},$$

the probability that h and c disagree on a randomly chosen input.

A hypothesis space H is said to be *trivial* if it consists of just one hypothesis, or if it consists of two hypotheses h and g such that

$$h(x) = 1 \Longleftrightarrow g(x) = 0.$$

That is, a hypothesis space is trivial if it consists of one hypothesis or if it consists of two hypotheses which are complementary. Trivial hypothesis spaces are not interesting, and we assume throughout that any hypothesis space we discuss is non-trivial.

Well-behaved hypothesis spaces and universal separability

In order to discuss the further measure-theoretic conditions to be placed on the hypothesis spaces, we need some definitions.

Let $0 < \epsilon < 1$ and $c \in H$, and denote by B_ϵ the set of hypotheses from H which have error greater than ϵ with respect to c. That is,

$$B_\epsilon = \{h \in H : \mathrm{er}_\mu(h,c) > \epsilon\}.$$

Given any positive integer m and any $\mathrm{x} = (x_1, x_2, \ldots, x_m) \in X^m$, let

$$\mathrm{er}_\mathrm{x}(h) = \frac{1}{m}\left|\{i : h(x_i) \neq c(x_i)\}\right|,$$

the *observed error* of h on x with respect to c.

We define two sets which will be crucial in our analysis.

Definition 1.3 *With m, c, μ, ϵ as above, define*

$$Q_m = Q_m^{\epsilon}(c, \mu) = \{\mathrm{x} \in X^m : \exists h \in B_\epsilon \text{ with } \mathrm{er}_{\mathrm{x}}(h) = 0\}.$$

Further, for a positive integer k and $0 < r < 1$, let

$$J_{m+k} = J_{m+k}^{\epsilon}(c, r, \mu) = \left\{\mathrm{xy} \in X^{m+k} : \exists h \in B_\epsilon \text{ s.t. } \mathrm{er}_{\mathrm{x}}(h) = 0,\ \mathrm{er}_{\mathrm{y}}(h) > r\epsilon\right\}.$$

□

It will become clear later why these sets are of interest. The analysis presented in Chapter 3 requires that for all m, Q_m be a measurable subset of X^m (with respect to the product σ-algebra Σ^m) and that for all m, k, J_{m+k} be a measurable subset of X^{m+k} (with respect to Σ^{m+k}). Further, this must be true for all possible c, ϵ, r and μ. Ben-David (see [11]) has called hypotheses spaces with this property *well-behaved.*

Definition 1.4 *The hypothesis space H is well-behaved if the sets Q_m and J_{m+k} defined above are measurable subsets of X^m and X^{m+k} (respectively) for all $c, \epsilon, m, k, r, \mu$.* □

This definition is an awkward one to have to work with. However, we can introduce a stronger restriction to place on H which will imply that H is well-behaved. This condition, known as *universal separability*, was introduced by Ben-David (see [11]).

Definition 1.5 *The hypothesis space H is universally separable if there is a countable subset H_0 of H such that any hypothesis in H is the pointwise limit of some sequence in H_0. In this case, we say that H is universally separable by H_0.* □

Thus, H is universally separable by H_0 if the following holds: Given $h \in H$ there is a sequence $(h_i)_{i=1}^{\infty}$ of hypotheses in H_0 such that for every $x \in X$, there is $n(x)$ for which

$$i \geq n(x) \Longrightarrow h_i(x) = h(x).$$

Ben-David proved (essentially) the following result. We omit the proof, but see Blumer *et al* [11].

Proposition 1.6 *If the hypothesis space H is universally separable then H is well-behaved* □

Thus there is an easily-described condition on H which ensures that H is well-behaved and therefore that all necessary sets are measurable.

1.5 Learnability: Further Discussion

The aim of this section is to again formally define learnability and to develop further the notation and terminology that shall be used in later chapters.

With the measurability details behind us, we shall now suppose that all sets we wish to measure are indeed measurable. This is the case if X and H are as described in the previous section.

Approximating the target concept by a hypothesis

Above, we gave a natural definition of how good an approximation a given hypothesis h is to the target concept c. The error $\mathrm{er}_\mu(h, c)$ of h with respect to c (and the underlying probability distribution μ on the input space) is defined to be the probability that h and c disagree on a randomly chosen input. We shall often use notation $\mathrm{er}_\mu(h)$ when c is clear from the context.

Given a target concept c from H and $\mathrm{x} = (x_1, x_2, \ldots, x_m) \in X^m$, we denote by $H[\mathrm{x}, c]$ the set of hypotheses from h which agree with c on $x_1, \ldots, x_m$ and we call this set the set of hypotheses *consistent* with c on x. Thus,

$$\begin{aligned} H[\mathrm{x}, c] &= \{h \in H : h(x_i) = c(x_i)\ (1 \leq i \leq m)\} \\ &= \{h \in H : \mathrm{er}_{\mathrm{x}}(h) = 0\}. \end{aligned}$$

When the target concept c is clear from the context, we shall denote $H[\mathrm{x}, c]$ simply by $H[\mathrm{x}]$ and describe it as the set of hypothesis consistent on or with x.

Given any subset F of H, we may define $\mathrm{haz}_\mu(F, c)$, the *haziness* of F, to be

$$\mathrm{haz}_\mu(F, c) = \sup\{\mathrm{er}_\mu(f) : f \in F\}.$$

Again, if c is clear, we use the notation $\mathrm{haz}_\mu(F)$. Thus, $\mathrm{haz}_\mu(F)$ is a measure of the worst error with respect to c and μ that any hypothesis from F can have.

Given that the learner chooses a hypothesis consistent with the target concept on the training sample, the definition of learnability is motivated by the requirement that the sample-size m is large enough so that the sample is representative of the target concept on the whole of the input space. That is, we wish to ensure that m is large enough so that there is a low probability that a training sample x of length m is "bad". We can formalise this using the notation developed in this section:

The sample is "bad" if there is some hypothesis from H which is consistent with the target on the sample but has error larger than ϵ (the desired accuracy) with respect to the target concept (and the probability distribution on the input space). Thus, the set of bad samples of length m is precisely the set

$$Q = \{\mathrm{x} \in X^m : B_\epsilon \cap H[\mathrm{x}, c] \neq \emptyset\},$$

defined earlier.

We wish the event Q to have a low probability. The probability referred to here is the natural product measure μ^m on the product σ-algebra Σ^m. Given a (small) real number δ strictly between 0 and 1, one could demand that

$$\mu^m(Q) < \delta.$$

As earlier, we call δ the *confidence parameter* and we call the quantity $1 - \delta$ the *confidence.*

We can now re-define learnability.

Definition 1.7 *The hypothesis space H, defined over the input space X is learnable if*

- *for every target concept $c \in H$,*
- *for every probability distribution μ on (the σ-algebra Σ of subsets of) X,*
- *for every accuracy parameter ϵ and confidence parameter δ, $(0 < \epsilon, \delta < 1)$,*

there is a sufficient sample-size $m_0 = m_0(\epsilon, \delta)$ such that

$$m \geq m_0 \Longrightarrow \mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}, c]) > \epsilon\} < \delta.$$

□

That is, H is learnable if there is a sufficient sample size $m_0 = m_0(\epsilon, \delta)$ such that given any target concept c from H and given any sample of $m \geq m_0$ inputs chosen according to *any* fixed distribution μ on X, if $h \in H$ is consistent with c on the sample then, with probability at least $1 - \delta$, h is an approximation to c with error less than ϵ.

In Definition 1.7, we usually omit explicit reference to any target concept c, since the defining property is required to hold for *all* c in H. Therefore, from now on, we shall write the last line of the definition as

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < \delta.$$

Notice that the m_0 of the definition does not depend in any way on μ or the target concept c. The learner knows neither the target concept nor the distribution and so must be able to use a sample of a size independent of these to be guaranteed good generalization (with fixed high probability).

1.6 Finite Hypothesis Spaces

The definition of learnability may at first sight look difficult to satisfy in general, because although it must hold for any probability distribution on the input space and for any target concept from the hypothesis space, the sufficient sample-size must be *independent* of both the distribution and the target. We show here that any finite hypothesis space is learnable. This is the easiest of the learnability results and can be regarded as "folklore". ~~The result is not too surprising; given a large enough sample, with high probability the sample contains most of the significant inputs (that is, those which have a high probability) and if a hypothesis h is found which agrees with the target concept on this sample, it should be a good approximation to c.~~

Indeed, suppose that H is a finite set of $\{0,1\}$-valued functions defined on an input space X and that $c \in H$. Let μ be any probability measure defined on X and suppose that $h \in H$ has error $\mathrm{er}_\mu(h) > \epsilon$ with respect to c and μ; that is, $h \in B_\epsilon$. Since $\mathrm{er}_\mu(h) > \epsilon$, we have

$$\mu\{x \in X : h(x) = c(x)\} = 1 - \mathrm{er}_\mu(h) < 1 - \epsilon.$$

Therefore

$$\mu^m\{\mathrm{x} \in X^m : h \in H[\mathrm{x}]\} < (1-\epsilon)^m.$$

This holds for any $h \in B_\epsilon$ and therefore

$$\begin{aligned}\mu^m\{\mathrm{x} \in X^m : B_\epsilon \cap H[\mathrm{x}] \neq \emptyset\} &= \mu^m\{\mathrm{x} \in X^m : \exists h \in B_\epsilon \text{ such that } h \in H[\mathrm{x}]\} \\ &\leq |B_\epsilon|\,(1-\epsilon)^m \\ &\leq |H|\,(1-\epsilon)^m.\end{aligned}$$

Now, for any $0 < \epsilon < 1$ and for any positive integer m,

$$(1-\epsilon)^m < \exp(-\epsilon m).$$

Therefore

$$\mu^m\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < |H|\exp(-\epsilon m).$$

Now,

$$|H|\exp(-\epsilon m) \leq \delta \iff m \geq \frac{1}{\epsilon}\log\left(\frac{|H|}{\delta}\right),$$

and so we have shown:

Theorem 1.8 *If H is a finite hypothesis space then H is learnable. A suitable value of $m_0(\epsilon, \delta)$ is*

$$m_0(\epsilon, \delta) = \left\lceil \frac{1}{\epsilon} \log \left(\frac{|H|}{\delta} \right) \right\rceil .$$

□

Notice that this proof relies very heavily on the finiteness of H and can in no way be extended to infinite H. That learnability can hold at all for infinite hypothesis spaces will be the subject of Chapter 3.

Chapter 2
The Vapnik-Chervonenkis Dimension

2.1 Introduction

We have seen that any finite hypothesis space is learnable. A key problem is to determine which infinite hypothesis spaces are learnable. To this end, one can make use of the *Vapnik-Chervonenkis dimension* or *VC dimension*, a combinatorial parameter associated with a hypothesis space. This parameter, introduced by Vapnik and Chervonenkis [36] and discussed in [19], can, in a sense, be regarded as a measure of the expressive power of the space.

2.2 The Vapnik-Chervonenkis Dimension

Shattering

Suppose that H is a collection of subsets of the set X. For any finite subset S of X, let $S \cap H$ be the collection

$$S \cap H = \{S \cap h : h \in H\}$$

of subsets of X. We shall call the sets of $S \cap H$ the *dichotomies of S by H*. We use this term as each $h \in H$ creates a dichotomy of S; a partition of S into two parts. The set S is said to be *shattered* by H if the set of dichotomies of S by H is the set of all subsets of S. Thus, S is shattered by H if every subset T of S can be expressed as

$$T = S \cap h,$$

for some $h \in H$.

If H is a set of $\{0,1\}$-valued functions, we say that a subset S of X is shattered by H if S is shattered by the collection

$$H^{-1}(1) = \left\{h^{-1}(1) : h \in H\right\}$$

of subsets of X. Alternatively, if $S = \{x_1, \ldots, x_m\}$, then S is shattered by H if the vectors

$$(h(x_1), \ldots, h(x_m))$$

yield all binary vectors of length m as h runs through H. We may also define the shattering of a vector in X^m. The vector $\mathrm{x} = (x_1, \ldots, x_m) \in X^m$ is shattered by H if $x_1, x_2, \ldots, x_m$ are distinct and if the set $\{x_1, \ldots, x_m\}$ is shattered by H; that is, if any binary vector b of length m can be expressed in the form

$$\mathrm{b} = (h(x_1), \ldots, h(x_m)),$$

for some $h \in H$.

The index function

For a collection of subsets H of a set X and for any finite subset S of X, the number of possible dichotomies of S by H is at most the number of distinct subsets of S; that is, 2^m where $m = |S|$. Further, S is shattered by H precisely when the number of such dichotomies is 2^m. Thus a useful quantity to measure is the number of dichotomies of S by H.

Fix the positive integer m, and for an m-subset S of X, let $\Pi_{m,H}(S)$ be the number of dichotomies of S by H. Thus

$$\Pi_{m,H}(S) = |\{S \cap h : h \in H\}| \leq 2^m.$$

This defines a function

$$\Pi_{m,H} : \binom{X}{m} \to \{1, 2, \ldots 2^m\}$$

from the set of all m-subsets of X to the integers between 1 and 2^m. S is shattered by H if and only if $\Pi_{m,H}(S) = 2^m$. These functions, one for each positive integer m, can be subsumed by defining the function

$$\Pi_H : \bigcup_{m=1}^{\infty} \binom{X}{m} \to \mathbf{N}$$

from the set of all finite subsets of X to the set of positive integers by

$$\Pi_H(S) = \Pi_{|S|,H}(S) \quad (S \subseteq X,\ |S| < \infty).$$

Thus S is shattered by H if and only if $\Pi_H(S) = 2^{|S|}$. $\Pi_H(S)$ is called the *index of S in H* and Π_H is the *index function (for H).*

Again, an analogous definition can be made for Boolean-valued functions defined on X. Fix a positive integer m, and for any $\mathrm{x} = (x_1, x_2, \ldots, x_m) \in X^m$, define the function

$$\mathbf{x}^* : H \to \{0,1\}^m$$

by

$$\mathbf{x}^*(h) = (h(x_1), \ldots, h(x_m)).$$

Then the image of H under $\mathbf{x}^*$ is the set of all binary vectors expressible in the form $(h(x_1), \ldots, h(x_m))$ for some $h \in H$. With this in mind, we define

$$\Pi_{m,H} : X^m \to \{1, 2, \ldots, 2^m\}$$

by

$$\Pi_{m,H}(\mathrm{x}) = |\mathbf{x}^*(H)| \quad (\mathrm{x} \in X^m).$$

We then define

$$\Pi_H : \bigcup_{m=1}^{\infty} X^m \to \mathbf{N}$$

by

$$\Pi_H(\mathrm{x}) = \Pi_{m,H}(\mathrm{x}) \quad (\mathrm{x} \in X^m).$$

Thus $\mathrm{x} \in X^m$ is shattered by H if and only if $\Pi_H(\mathrm{x}) = 2^m$. Again, $\Pi_H(\mathrm{x})$ is called the *index of* x *in H* and Π_H is the *index function (for H).*

We remark that when H is a set of Boolean-valued functions defined on X, for a finite subset S of X, $\Pi_H(S)$ is the cardinality of

$$H|S = \{h|S : h \in H\},$$

the set of functions in H restricted to domain S.

The growth function and the VC dimension

The *Vapnik-Chervonenkis dimension*, or *VC dimension*, of a family H of subsets of a set X is defined to be the supremum of the cardinalities of the subsets of X shattered by H. The definition allows for H to have infinite VC dimension if for each positive integer m, there is an m-subset of X shattered by H.

For a family H of boolean-valued functions defined on X, the VC dimension is infinite if given any positive integer m, there is some $\mathrm{x} = (x_1, x_2, \ldots, x_m) \in X^m$ such that every binary vector b of length m can be expressed as

$$\mathrm{b} = (h(x_1), \ldots, h(x_m))$$

for some $h \in H$. Otherwise, the VC dimension is the largest m for which this holds.

If H is a collection of subsets of a set X or a collection of boolean-valued functions defined on X, we define the *growth function (of H)*

$$\Pi_H : \mathbf{N} \to \mathbf{N}$$

by

$$\Pi_H(m) = \max\{\Pi_H(\mathrm{x}) : \mathrm{x} \in X^m\} = \sup \Pi_{m,H}.$$

The same notation is used for the growth function and the index function, but this should cause no confusion. Then the VC dimension of H is

$$\mathrm{VCdim}(H) = \sup\{m : \Pi_H(m) = 2^m\},$$

where we take the supremum to be infinity if this set is unbounded. Clearly, if H is a set of subsets of X and there is no m-subset of X shattered by H then there is no $(m+1)$-subset of X shattered by H. An analogous observation holds for the case in which H is a set of functions. Therefore, the VC dimension of H is either infinity or is the least integer d such that

$$\Pi_H(d) = 2^d, \quad \Pi_H(d+1) \neq 2^{d+1}.$$

The following elementary observation (made, for example in [19]) applies when H is finite.

Proposition 2.1 *Let H be a finite set of subsets of a set X (or, equivalently, of Boolean-valued functions on X). Then H has a finite VC dimension of at most $\log_2 |H|$.*

Proof If H shatters an s-subset of X, then there are at least 2^s distinct sets (or functions) in H, at least one for each dichotomy of the shattered set. Therefore

$$2^s \leq |H|,$$

and hence

$$s \leq \log_2 |H|.$$

The result follows. □

The following observation will prove useful later in this chapter.

Proposition 2.2 *If H is a set of subsets of a finite set X (or, equivalently, of Boolean-valued functions defined on X), then*

$$|H| = \Pi_H(|X|).$$

Proof Two members of H are distinct if and only if they induce distinct dichotomies of X, and therefore $|H|$ is the number of dichotomies of X by H. That is,

$$|H| = \Pi_H(X).$$

The result follows since, clearly, $\Pi_H(X) = \Pi_H(|X|)$. □

Sauer's lemma

As one might suspect, the growth function for a set H of sets or Boolean-valued functions can be related to the VC dimension of H. We know that if H has infinite VC dimension then, by definition, for all positive integers m, $\Pi_H(m) = 2^m$. When H has finite VC dimension d then for all m less than or equal to d, $\Pi_H(m) = 2^m$, and for all m greater than d, $\Pi_H(m) < 2^m$. An interesting and useful result is that, although the values of $\Pi_H(m)$ for $m \leq d$ are the values of the exponential function 2^m, the growth function is actually bounded by a polynomial function of m.

Before proving this result, known as Sauer's Lemma, in the form we require, some preliminaries are needed. (Actually, the result we seek is a corollary of a result of Sauer, but we will refer to it as Sauer's Lemma).

Following Vapnik and Chervonenkis [36], for $d, m \geq 1$, we shall denote by $\Phi(d, m)$ the maximum number of components or cells into which it is possible to partition d-dimensional Euclidean space by means of m hyperplanes.

It can be shown that

$$\Phi(d, m) = \begin{cases} 2^m & \text{if } m \leq d \\ \sum_{k=0}^{d} \binom{m}{k} & \text{if } m > d \end{cases}.$$

We extend the definition of Φ according to this formula, defining $\Phi(0, m) = 1$ for all $m \geq 0$ and $\Phi(d, 0) = 1$ for all $d \geq 0$. An important observation is that the function Φ satisfies the relation

$$\Phi(d, m) = \Phi(d, m-1) + \Phi(d-1, m-1).$$

The following result is essentially due to Sauer [29] and our proof is similar to that in [19].

Theorem 2.3 *If H is a hypothesis space of finite VC dimension d then for all positive integers m,*

$$\Pi_H(m) \leq \Phi(d, m).$$

In particular, for $m \geq d$,

$$\Pi_H(m) \leq \sum_{k=0}^{d} \binom{m}{k}.$$

Proof We prove the result for the case in which H is a set of subsets of the space X. (The result for H a set of Boolean-valued functions defined on X follows immediately from this).

The result clearly holds true for $d = 0$, for in this case, for any positive integer m, both sides of the inequality are equal to 1. The result also clearly holds when $m = 1$ and $d \geq 1$, for we have $\Pi_H(1) \leq 2 = \Phi(d, 1)$. Assume therefore that $m \geq 0$, $d \geq 0$. Assume also, inductively, that if G is any hypothesis space of VC dimension at most d, then $\Pi_G(m) \leq \Phi(d, m)$ and that if G is any hypothesis space of VC dimension at most $d + 1$ then $\Pi_G(m) \leq \Phi(d + 1, m)$. Now suppose that H is a hypothesis space of subsets of X and that H has VC dimension at most $d + 1$. Let $S \subseteq X$ be an $(m + 1)$-subset of X and let

$$H_S = S \cap H = \{S \cap h : h \in H\}$$

be the set of dichotomies of S induced by H. Clearly, by definition,

$$|H_S| = \Pi_H(S).$$

Choose $x \in S$, and let

$$H_S - x = \{T \setminus \{x\} : T \in H_S\} = \{(S \cap h) \setminus \{x\} : h \in H\}$$

and

$$\overline{H}_S = \{T \in H_S : x \notin T, T \cup \{x\} \in H_S\}.$$

Then

$$|H_S| = |H_S - x| + \left|\overline{H}_S\right|.$$

To see this, observe that, under the mapping which removes x from the sets of H_S, any two sets of the form T and $T \cup \{x\}$ map to the same set. That is,

$$\begin{aligned} |H_S - x| &= |\{T \setminus \{x\} : T \in H_S\}| \\ &= |H_S| - |\{U \in H_S : x \notin U, U = T \setminus \{x\} \text{ for some } T \in H_S\}| \\ &= |H_S| - \left|\overline{H}_S\right|. \end{aligned}$$

Now,

$$\begin{aligned} |H_S - x| &= |\{(S \cap h) \setminus \{x\} : h \in H\}| \\ &= |\{(S \setminus \{x\}) \cap h : h \in H\}| \\ &= \Pi_H(S \setminus \{x\}) \\ &\leq \Phi(d+1, m), \end{aligned}$$

by induction, since $S \setminus \{x\}$ is an m-set. Now consider

$$\overline{H}_S = \left\{\overline{h} \in H_S : x \notin \overline{h}, \overline{h} \cup \{x\} \in H_S\right\} \subseteq H_S$$

as a hypothesis space of subsets of $S \setminus \{x\}$. Then, by Proposition 2.2,

$$\left|\overline{H}_S\right| = \Pi_{\overline{H}_S}(S \setminus \{x\}).$$

We claim that the hypothesis space $\overline{H}_S$ has VC dimension at most d. To see this, suppose that $R \subseteq S \setminus \{x\}$ is shattered by $\overline{H}_S$. Clearly, $x \notin R$. For any subset A of R, there is $\overline{h} \in \overline{H}_S$ such that $A = R \cap \overline{h}$. Now, $x \notin A$, $x \notin \overline{h}$ and $\overline{h} \cup \{x\} \in H_S$, (by definition of $\overline{H}_S$). Therefore,

$$A = (R \cup \{x\}) \cap \overline{h}$$

and

$$A \cup \{x\} = (R \cup \{x\}) \cap \left(\overline{h} \cup \{x\}\right).$$

It follows that the set $R \cup \{x\}$ is shattered by H_S and $|R \cup \{x\}| \leq d+1$, since H_S has VC dimension at most $d+1$. Hence $|R| \leq d$, and so $\overline{H}_S$ has VC dimension at most d. By induction,

$$\left|\overline{H}_S\right| = \Pi_{\overline{H}_S}(S \setminus \{x\}) \leq \Phi(d, m).$$

Therefore,

$$\begin{aligned}\Pi_H(S) &= |H_S - x| + \left|\overline{H}_S\right| \\ &\leq \Phi(d+1, m) + \Phi(d, m) \\ &= \Phi(d+1, m+1),\end{aligned}$$

and the result follows. □

The result we seek is a corollary of this, and the following result from [11] takes us some way towards it.

Proposition 2.4 *For $m \geq d \geq 1$, we have*

$$\Phi(d, m) \leq \frac{2m^d}{d!}.$$

Proof If $d = 1$ then

$$\Phi(d, m) = m + 1 \leq 2m.$$

If $m = d > 1$ then $\Phi(d, m) = 2^d$. Now, for $d \geq 1$, we have

$$\left(1 + \frac{1}{d}\right)^d \geq 1 + d\frac{1}{d} = 2,$$

and therefore, making the obvious inductive hypothesis,

$$\begin{aligned}2^{d+1} &\leq \left(\frac{d+1}{d}\right)^d 2^d \\ &\leq 2\left(\frac{d+1}{d}\right)^d \frac{d^d}{d!} \\ &= 2\frac{(d+1)^d}{(d+1)!},\end{aligned}$$

verifying the result for $m = d > 1$.

Suppose that $m > d \geq 1$. Now,

$$\Phi(d+1, m+1) = \Phi(d+1, m) + \Phi(d, m),$$

so it suffices to prove that

$$2\frac{m^{d+1}}{(d+1)!} + 2\frac{m^d}{d!} \leq 2\frac{(m+1)^{d+1}}{(d+1)!}.$$

This is true if and only if

$$\begin{aligned} & m^{d+1} + (1+d)m^d \leq (m+1)^{d+1} \\ \iff & (d+m+1)m^d \leq (m+1)^{d+1} \\ \iff & \frac{d+m+1}{m} \leq \frac{(m+1)^{d+1}}{m^{d+1}} \\ \iff & 1 + \left(\frac{d+1}{m}\right) \leq \left(1+\frac{1}{m}\right)^{d+1}, \end{aligned}$$

and this last inequality follows from the binomial theorem. □

We shall call the following result Sauer's Lemma.

Theorem 2.5 [Sauer's Lemma] *If H is a hypothesis space of finite VC dimension $d \geq 1$ then for all positive integers $m \geq d$, we have*

$$\Pi_H(m) < \left(\frac{em}{d}\right)^d.$$

Proof In view of the preceeding two results, we have, for $m \geq d$,

$$\Pi_H(m) \leq \Phi(d,m) \leq 2\frac{m^d}{d!},$$

and therefore it suffices to show that for all $m \geq d \geq 1$,

$$2\left(\frac{d}{e}\right)^d < d!$$

This can be proved using Stirling's Approximation, but we shall give a simple proof by induction on d. The result clearly holds when $d = 1$. Making the inductive hypothesis, for $d \geq 1$ we have

$$(d+1)! = (d+1)\,d! \geq (d+1)\,2\left(\frac{d}{e}\right)^d.$$

It suffices to prove that

$$(d+1)\,2\left(\frac{d}{e}\right)^d > 2\left(\frac{d+1}{e}\right)^{d+1}.$$

This is true if and only if

$$\left(1+\frac{1}{d}\right)^d < e,$$

which is true for any $d \geq 1$. The result follows. □

This last proposition shows that the function $\Pi_H(m)$ is bounded by a polynomial in m of degree d, where d is the VC dimension of H.

2.3 Examples

Rays on the real line

We start with a very simple example to illustrate the ideas of this chapter. A *ray* on the real line is an interval of the form $(-\infty, a]$ for some $a \in \mathbf{R}$. It is trivial to see that if x and y are real numbers satisfying $x < y$, then one cannot find a real number a such that $y \in (-\infty, a]$ and $x \notin (-\infty, a]$. That is, if X is taken to be the real line and H is the set of all rays on the real line, then

$$\mathrm{VCdim}(H) < 2.$$

But H certainly shatters any singleton subset of X and therefore H has VC dimension 1.

Let

$$x_1 < x_2 < \ldots < x_m$$

be any m distinct real numbers and let

$$S = \{x_1, x_2, \ldots, x_m\}.$$

Then the dichotomies of S by H are the empty set and the sets

$$\{x_1, \ldots, x_k\}, \quad (1 \leq k \leq m).$$

That is $\Pi_H(S) = m + 1$. Clearly, $\Pi_H(m) = m + 1$ since any set of m distinct points on the real line has the same number of dichotomies by H. Compare this with Theorem 2.3, which gives the result

$$\Pi_H(m) \leq 1 + m.$$

Therefore we can obtain equality in this theorem, and the bound of the theorem is tight. Later, we exhibit a general family of hypothesis spaces for which equality is achieved in Theorem 2.3.

Half-spaces of Euclidean space

If we allow the H of the previous example to additionally contain all intervals of the form $[b, \infty)$ for real numbers b, then the VC dimension of the resulting space H_1 is 2. This hypothesis space can be described as the set of all closed half-spaces of 1-dimensional Euclidean space.

Consider now the set H_2 of closed half-spaces of $\mathbf{R}^2$, and suppose that S is a set of four points in $\mathbf{R}^2$. Observe that $T \subseteq S$ is a dichotomy of S by H_2 if and only if T and $S \setminus T$ can be separated by a hyperplane (that is, a line). If any three of the points of S are collinear then the two of these points farthest from each other are not separable from the middle one by a hyperplane (line), and therefore not all subsets of S can be obtained as dichotomies of S by H. So suppose then that no three points in S are collinear. Then there are two possibilities to consider; either all four points lie on the boundary of the convex hull of S, or there is one point lying in the interior of the convex hull of S. In the first case, two opposite points cannot be separated from the other two points by a hyperplane, while in the second case, the point lying inside the convex hull is not separable from the other three points by a hyperplane. In all cases, then, S is not shattered by the set of half-spaces and therefore H_2 has VC dimension at most 3. But it has VC dimension at least 3 since any (non-degenerate) triangle of points can be shattered.

More generally, we have

Theorem 2.6 *If H_n is the set of half-spaces of $\mathbf{R}^n$ then H_n has VC dimension $n + 1$.*

Proof Firstly, we observe that for a finite set S of points of $\mathbf{R}^n$, the subset T of S is a dichotomy of S by H_n if and only if T and $S \setminus T$ lie in different open half-spaces on either side of some hyperplane of $\mathbf{R}^n$. That is, T is a dichotomy of S by H_n if and only if there is some hyperplane such that the points of T lie strictly on one side of the hyperplane, and the points of $S \setminus T$

lie strictly on the other side of the hyperplane. (Because S is finite, we can insist no points of T or $S \setminus T$ lie on the hyperplane). We shall say that T and $S \setminus T$ are *linearly separable* if this condition holds. Now, open half-spaces are convex subsets of $\mathbf{R}^n$. It follows that if T and $S \setminus T$ are linearly separable by the hyperplane L, then the convex hulls of T and $S \setminus T$ lie strictly on different sides of L and therefore have no points in common. (We interpret the convex hull of the empty set as the empty set). Conversely, if $T \subseteq S$ and the convex hulls of T and $S \setminus T$ do not intersect then T is a dichotomy of S by H_n. Now, Radon's theorem [15] asserts that if S is any set of $n+2$ points in n-dimensional Euclidean space, then there is a partition of S into two non-empty disjoint subsets S_1 and S_2 such that the convex hulls of S_1 and S_2 intersect. It follows that S_1 and S_2 are not linearly separable and therefore that S_1 and S_2 are not dichotomies of S by H_n. Therefore H_n has VC dimension at most $n+1$.

Conversely, let o denote the origin and, for $1 \leq i \leq n$, let e_i be the point of $\mathbf{R}^n$ with a 1 in position i and every other entry 0. Then it is easy to see that for any $T \subseteq S = \{\mathrm{o}, \mathrm{e}_1, \ldots, \mathrm{e}_n\}$, the convex hulls of T and $S \setminus T$ have empty intersection. Therefore, S is shattered by H_n. Consequently, the VC dimension of H_n is at least $n+1$, and the theorem follows. □

We now show that Theorem 2.3 is tight, and that equality is achieved for some hypothesis space of each finite VC dimension.

For any n, let G_n be the set of all subsets of $\mathbf{R}^n$ of the form

$$g_{\mathrm{y}} = \{\mathrm{x} \in \mathbf{R}^n : \langle \mathrm{x}, \mathrm{y} \rangle \geq 1\},$$

where $\langle \mathrm{x}, \mathrm{y} \rangle$ denotes the inner product of x and y. We shall call the space G_n the space of *one-side half-spaces* of $\mathbf{R}^n$. Observe that the members of G_n are precisely the closed half-spaces of $\mathbf{R}^n$ not containing the origin. We have

Theorem 2.7 *Let G_n be the set of all one-sided half-spaces of $\mathbf{R}^n$, as defined above. Then for any positive integer m,*

$$\Pi_{G_n}(m) = \Phi(n, m).$$

Proof The proof of this is as in [36]. Given any $\mathrm{x} \in \mathbf{R}^n$, there is a partition of G into those g_{y} such that $\langle \mathrm{x}, \mathrm{y} \rangle \geq 1$ and those g_{y} such that $\langle \mathrm{x}, \mathrm{y} \rangle < 1$, and an obvious corresponding partition of $\mathbf{R}^n$. Let

$$S = \{\mathrm{x}_1, \ldots, \mathrm{x}_m\}$$

be any m-subset of $\mathbf{R}^n$. Then from S we obtain a partition of $\mathbf{R}^n$ by m hyperplanes into a number of components or cells such that for all vectors y belonging to one particular cell, g_{y} induces a particular dichotomy of S, and this differs from the dichotomies induced by g_{z} for z in any other cell. Thus the number of dichotomies of S by G is the number of distinct cells obtained, and $\Pi_G(m)$ is therefore the maximum number of components into which it is possible to partition n-dimensional Euclidean space by means of m hyperplanes. By definition, this is $\Phi(n, m)$. □

We now show

Theorem 2.8 *With G_n as above, G_n has VC dimension n.*

Proof Suppose that S is any set of $n+1$ points in $\mathbf{R}^n$ and suppose, with the view to obtaining a contradiction, that S is shattered by $G = G_n$. Observe that the origin o cannot belong to any set in G, since for any y, the inner product of o and y is 0, which is less than 1. Therefore, the origin is not one of the points of S. Now consider the set $S^* = S \cup \{\mathrm{o}\}$. This is an $(n+2)$-subset of $\mathbf{R}^n$, and so, by Radon's theorem, has a partition into two subsets S_1^* and S_2^* such that the convex hulls of S_1^* and S_2^* intersect. Now, one of S_1^*, S_2^* contains the origin; without loss of generality we suppose this is S_2^*. Because the convex

hulls of S_1^* and S_2^* intersect, there is no closed half-space h of $\mathbf{R}^n$ such that $S^* \cap h = S_1^*$. Let

$$S_1 = S_1^* \setminus \{o\} = S_1^*, \quad S_2 = S_2^* \setminus \{o\}.$$

If, for some $g \in G$, we have $S_1 = S \cap g$, then

$$\begin{aligned} S^* \cap g &= (S_1^* \cup S_2^*) \cap g \\ &= (S_1 \cap g) \cup ((S \setminus S_1) \cap g) \cup (\{o\} \cap g) \\ &= S_1 = S_1^*, \end{aligned}$$

since $o \notin g$. This contradicts the above. It follows that the set S is not shattered by G and, consequently, G has VC dimension at most n.

Conversely, let the points $e_1, e_2, \ldots, e_n$ be as before, and let

$$S = \{e_1, e_2, \ldots, e_n\}.$$

Then we claim that S can be shattered by G. Indeed, consider $S^* = S \cup \{o\}$. We saw in the course of proving the previous result that S^* is shattered by the space of closed half-spaces of $\mathbf{R}^n$. Let T be any subset of S, and let $T^* = T \cup \{o\}$. Then there is $h_1 \in H_n$ such that $T = S^* \cap h_1$ and there is $h_2 \in H_n$ such that $T^* = S^* \cap h_2$. Now, h_1 is a closed half-space of $\mathbf{R}^n$ and $o \notin h_1$, so $h_1 \in G$. Therefore there is $g_1 = h_1 \in G$ such that $T = S^* \cap g_1$. Further, since T^* is a dichotomy of S^* by H_n, there is a hyperplane L strictly separating T^* from $S^* \setminus T^* = S \setminus T$. Therefore there is $g_2 \in G$ such that $S \setminus T = S \cap g_2$. (We take g_2 to be that half-space determined by L which does not contain the origin). It follows that S is shattered by G and the VC dimension of G_n is at least n. The theorem follows. □

Theorem 2.3 implies, since G_n has VC dimension n, that $\Pi_{G_n}(m)$ is at most $\Phi(n, m)$. Therefore, the inequality of the theorem becomes an equality for the spaces G_n, and the theorem is tight.

A *positive half-space* of $\mathbf{R}^n$ is a set of the form

$$p_y = \{x \in \mathbf{R}^n : \langle x, y \rangle \geq 0\},$$

for some $y \in \mathbf{R}^n$. We remark that in a manner similar to that of the proof of Theorem 2.8, one can easily show that the set of all positive half-spaces of $\mathbf{R}^n$ has VC dimension n.

Boolean threshold functions

A Boolean function

$$f : \{0,1\}^n \to \{0,1\}$$

is called a *Boolean threshold function* or *Boolean linear threshold function* if there is a vector

$$\mathrm{w} = (w_1, w_2, \ldots, w_n) \in \mathbf{R}^n$$

and a constant $\theta \in \mathbf{R}$ such that

$$f(x_1, x_2, \ldots, x_n) = 1 \iff w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \geq \theta.$$

That is, f returns a 1 if the weighted sum of the inputs exceeds or equals a certain threshold, and returns a 0 otherwise. Geometrically, a Boolean threshold function is determined by a closed half-space of $\mathbf{R}^n$; the inputs for which f computes 1 are those vertices of the n-dimensional unit cube which lie on the side

$$\{\mathrm{x} : \langle \mathrm{x}, \mathrm{w} \rangle \geq \theta\}$$

of the hyperplane

$$\{\mathrm{x} : \langle \mathrm{x}, \mathrm{w} \rangle = \theta\}.$$

Thus the space T_n of Boolean threshold functions on n variables can be described as the space of all closed half-spaces of $\mathbf{R}^n$, restricted to the unit cube. Therefore, the VC dimension of T_n, a restriction of a space of VC dimension $n+1$, is at most $n+1$.

But the set

$$S = \{\mathrm{o}, \mathrm{e}_1, \ldots, \mathrm{e}_n\}$$

described earlier is a subset of the set of vertices of the unit cube, and is shattered by the space of closed half-spaces of $\mathbf{R}^n$. It follows that T_n has VC dimension at least $n+1$. Therefore the VC dimension of T_n is $n+1$.

An interesting application of the VC dimension and Sauer's Lemma is to use these results to bound the number of Boolean threshold functions on n variables. For a simple lower bound, note that since $T = T_n$ has VC dimension $n+1$, it follows from Proposition 2.1 that

$$|T_n| \geq 2^{n+1}.$$

Muroga [24] has shown that

$$|T_n| \leq 2^{n^2}.$$

Using the powerful machinery described in this chapter, we can produce a significantly better upper bound.

Theorem 2.9 *The set T_n of Boolean threshold functions defined on $\{0,1\}^n$ satisfies*

$$|T_n| = O\left(2^{n^2+3n-(n+1)\log_2(n+1)}\right).$$

Proof Let $X = \{0,1\}^n$ be the input space to $T = T_n$. The VC dimension of T is $n+1$ and therefore, by Proposition 2.2, we have

$$\begin{aligned}
|T| &= \Pi_T(|X|) \\
&< \left(\frac{e\,|X|}{n+1}\right)^{n+1} \\
&= \left(\frac{e2^n}{n+1}\right)^{n+1} \\
&= \left(\frac{e}{n+1}\right)\left(\frac{2e}{n+1}\right)^n 2^{n^2} \\
&= O\left(2^{n^2+3n-(n+1)\log_2(n+1)}\right).
\end{aligned}$$

□

Thus $|T_n|$ is significantly less that 2^{n^2}. Compare this number with the total number of Boolean functions of n variables, which is 2^{2^n}.

Graph neighbourhoods

Following Haussler and Welzl [19] we may, as a further example on the VC dimension, define the VC dimension of a graph. Let $G = (V, E)$ be a (simple, loopless) graph with vertex-set V and edge-set E. The *neighbourhood* of a vertex v is the set

$$N(v) = \{u \in V : \{u, v\} \in E\} \cup \{v\},$$

the set of all vertices at distance at most 1 from v. Denote by $N(G)$ the set of all neighbourhoods of vertices of G,

$$N(G) = \{N(v) : v \in V\}.$$

Then $N(G)$, as a set of subsets of the set V, has a VC dimension, which we shall call the VC dimension of the graph G.

A graph G is said to be *homeomorphic* to a graph H if (an isomorphic copy of) G can be obtained from H by the addition and removal of vertices of degree two (the incidence being changed in the obvious manner). Further, a *subgraph* H of the graph $G = (V, E)$ is a graph of the form $H = (V_1, E_1)$, where $V_1 \subseteq V$ and $E_1 \subseteq E$. We now show the following.

Theorem 2.10 *If the graph G has VC dimension at least n, then G must contain a subgraph homeomorphic to the complete graph K_n on n vertices.*

Proof Suppose that S is a set of n vertices of G shattered by $N(G)$ and let x, y be any two vertices in S and suppose that x and y are not adjacent in G. Since S is shattered, the set $\{x, y\}$ can be obtained as a dichotomy of S by $N(G)$. Thus, there is a vertex $w = w(x, y)$ such that

$$S \cap N(w) = \{x, y\}.$$

If w is one of x or y, say x, then the above condition implies that $y \in N(w) = N(x)$. That is, if w is one of x, y, the above condition implies that x and y are

adjacent in G. Thus, w is neither x nor y and so w is some vertex in $V \setminus S$ such that the only vertices of S adjacent to w are x and y. This analysis holds for each pair of non-adjacent vertices in S. Let

$$V_1 = S \cup \{w(x,y) : x,y \in S,\ \ x,y \text{ not adjacent in } G\}$$

where, for any non-adjacent pair x,y of vertices from S, $w(x,y)$ is, as above, any vertex of G such that $N(w(x,y)) \cap S = \{x,y\}$. Further, let E_1 be the set of edges of G joining two vertices of S or a vertex of S and a vertex $w(x,y)$ of V_1. Then the subgraph $H = (V_1, E_1)$ of G is homeomorphic to the complete graph on n vertices; in H, any two vertices of S are adjacent or there is a vertex of degree two in H adjacent to each of the two vertices. The result follows. □

Note that it was not necessary for this result to have every subset of S equal to a dichotomy of S by $N(G)$; rather, all we required was that every 2-subset of S be a dichotomy of S by $N(G)$.

Chapter 3
Bounding Sample Size with the VC Dimension

3.1 Introduction

We have already seen that any finite hypothesis space is learnable, but it remains to consider the learnability of infinite hypothesis spaces. In this chapter, we show how to relate the VC dimension of a hypothesis space to the learnability of the space and to the sample-sizes sufficient for and necessary for the learnability of the space to given accuracy with a given confidence. We first give a new estimate of the probability of a bad training sample, involving the expectations of the index functions rather than the growth functions. A new sufficient sample-size for learning in a space with finite VC dimension is given, improving the best known previous bounds. A proof of a simple lower bound on necessary sample-size in terms of the VC dimension of the hypothesis space is presented, and we use this to show that if a hypothesis space is learnable, then it necessarily has finite VC dimension. We also present lower bound results of Blumer *et al* [11] and Ehrenfeucht *et al* [13]. The results of this chapter combine to give a key result in computational learning theory, due to Blumer, Ehrenfeucht, Haussler and Warmuth [11], which states that a hypothesis space is learnable if and only if it has finite VC dimension.

3.2 Bounding the Probability of a Bad Training Sample

Definitions

Recall the conditions placed on X and H in Chapter 1. We assume throughout that H is a non-trivial well-behaved hypothesis space defined on an input space X and that Σ is a σ-algebra of subsets of X which will be the power set of X if X is countable and will be the induced Borel σ-algebra if X is a subset of Euclidean space. Further, we assume that m, k are positive integers, $c \in H$ is some target concept, ϵ, r are real numbers strictly between 0 and 1 and μ is a measure such that (X, Σ, μ) is a probability space. We denote by B_ϵ the set

$$B_\epsilon = \{h \in H : \mathrm{er}_\mu(h) > \epsilon\}$$

of hypotheses from H which have error at least ϵ with respect to target concept c. Q denotes the set

$$Q = Q_\epsilon^m(c, \mu) = \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\},$$

and J denotes

$$J = J_\epsilon^{m+k}(c, r, \mu) = \left\{\mathrm{xy} \in X^{m+k} : \exists h \in B_\epsilon \,\mathrm{s.t.}\, \mathrm{er}_\mathrm{x}(h) = 0,\, \mathrm{er}_\mathrm{y}(h) > r\epsilon\right\}.$$

Notice that Q could equally well be described as

$$Q = \{\mathrm{x} \in X^m : \exists h \in H[\mathrm{x}] \;\text{ s.t. }\; \mathrm{er}_\mu(h) > \epsilon\}.$$

Since H is a well-behaved hypothesis space, for any values of r, ϵ, m, k, for any $c \in H$ and for any probability measure μ on (X, Σ), Q and J will be measurable; that is, they belong to the product σ-algebras Σ^m and Σ^{m+k} (respectively).

Measurability of the index function

Part of our main bounding theorem involves the expected value (or expectation) of the index function, when this expected value exists. For any bounded function, the expectation of the function exists if and only if the function is measurable. We have seen that if H is universally separable then it is well-behaved. We now show that if H is universally separable then the index functions Π_{m,B_ϵ} and $\Pi_{m,H}$ are Σ^m-measurable.

We first need the following result, in which for

$$\mathrm{y} = (y_1, y_2, \ldots, y_m) \in \{0,1\}^m$$

and $h \in H$, we define

$$h^{-1}(\mathrm{y}) = \{(x_1, x_2, \ldots, x_m) \in X^m : h(x_i) = y_i \, (1 \leq i \leq m)\}.$$

Lemma 3.1 *Suppose that H is a universally separable hypothesis space defined on an input space X and that H_0 is as in the definition of universal separability. Then, for any $\mathrm{y} \in \{0,1\}^m$, we have*

$$\bigcup_{h \in H} h^{-1}(\mathrm{y}) = \bigcup_{h \in H_0} h^{-1}(\mathrm{y}) \in \Sigma^m.$$

Proof Suppose that

$$\mathrm{y} = (y_1, y_2, \ldots, y_m)$$

and

$$\mathrm{x} = (x_1, x_2, \ldots, x_m) \in \bigcup_{h \in H} h^{-1}(\mathrm{y}).$$

Then there is $h \in H$ such that

$$h(x_i) = y_i \;\; (1 \leq i \leq m).$$

By universal separability of H by H_0, there is a sequence $(h_i)_{i=1}^{\infty}$ of hypotheses in H_0 such that h is the pointwise limit of this sequence. Therefore, for each $1 \leq k \leq m$, there is $n(k)$ such that

$$i \geq n(k) \Longrightarrow h_i(x_k) = h(x_k) = y_k.$$

Thus, for

$$n = \max\{n(k) : 1 \leq i \leq k\},$$

we have

$$\mathrm{x} \in h_n^{-1}(\mathrm{y}).$$

Hence

$$\bigcup_{h \in H} h^{-1}(\mathrm{y}) \subseteq \bigcup_{h \in H_0} h^{-1}(\mathrm{y}),$$

and the reverse containment is obvious. Now,

$$h^{-1}(\mathrm{y}) = h^{-1}(\{y_1\}) \times h^{-1}(\{y_2\}) \times \ldots \times h^{-1}(\{y_m\}) \in \Sigma^m,$$

and so

$$\bigcup_{h \in H} h^{-1}(\mathrm{y}),$$

as a countable union of measurable subsets of X^m, is measurable. □

This has the following implication.

Proposition 3.2 *If H is a universally separable hypothesis space over X then for each positive integer m, $\Pi_{m,H}$ is a Σ^m-measurable function.*

Proof Fix $\mathrm{y} \in \{0,1\}^m$ and let

$$H^{-1}(\mathrm{y}) = \bigcup_{h \in H} h^{-1}(\mathrm{y}) \subseteq X^m.$$

By the previous result, $H^{-1}(\mathrm{y})$ is a Σ^m-measurable subset of X^m. Now, for $\mathrm{x} \in X^m$,

$$\Pi_{m,H}(\mathrm{x}) = \sum_{\mathrm{y}} I_{H^{-1}(\mathrm{y})}(\mathrm{x}),$$

where the summation is over all $\mathrm{y} \in \{0,1\}^m$ and where $I_{H^{-1}(\mathrm{y})}$ is the characteristic (or indicator) function of $H^{-1}(\mathrm{y})$. It follows that $\Pi_{m,H}$ is a measurable function. □

In the same way, if H is universally separable then Π_{B_ϵ} is measurable for any $\epsilon > 0$.

Statement of the bounding theorem

For any positive integer n and for any real-valued Σ^n-measurable function ϕ defined on X^n, we shall denote by $\mathbf{E}(\phi(\mathrm{x}))$ the expected value, or expectation, with respect to μ^n of ϕ (over X^n).

In particular, for any hypothesis space F, when $\Pi_{n,F}$ is measurable, we denote by $\mathbf{E}(\Pi_{n,F}(\mathrm{x}))$ the expected value of $\Pi_{n,F}$. We shall omit the subcript n in what follows when it is clear from the context.

We are now in a position to state the main bounding theorem.

Theorem 3.3 *Let H be a hypothesis space of functions from an input space X to $\{0,1\}$. Let $0 < \epsilon < 1$ and m a positive integer. Suppose that μ is any probability measure on X and that $c \in H$ is any target concept. Let*

$$Q = \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\}.$$

Denote by B_ϵ the subset of H of hypotheses h for which $\mathrm{er}_\mu(h) > \epsilon$. For any positive integer $k > 1/\epsilon$ and for any r such that

$$0 < r < 1 - \frac{1}{\sqrt{\epsilon k}},$$

let the constant $C(r,k)$ be defined as

$$C(r,k) = \frac{\epsilon k(1-r)^2}{\epsilon k(1-r)^2 - 1}.$$

Then if Π_{B_ϵ} and Π_H are Σ^{m+k}-measurable functions (in particular, if H is universally separable), we have

$$\begin{aligned}\mu^m(Q) &< C(r,k)\,\mathbf{E}(\Pi_{B_\epsilon}(\mathrm{x}))\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1} \\ &\leq C(r,k)\,\mathbf{E}(\Pi_H(\mathrm{x}))\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1}.\end{aligned}$$

In any case,

$$\begin{aligned}\mu^m(Q) &< C(r,k)\,\Pi_{B_\epsilon}(m+k)\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1} \\ &\leq C(r,k)\,\Pi_H(m+k)\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1}.\end{aligned}$$

Here, $\mathbf{E}(.)$ *denotes expected value over* X^{m+k} *with respect to the product probability measure* μ^{m+k}. □

It is worth remarking that even when Π_H is not measurable, we may replace $\mathbf{E}(\Pi_H(\mathrm{x}))$ by

$$\inf \mathbf{E}(\phi(\mathrm{x})),$$

where the infimum is over all measurable functions ϕ which bound Π_H from above. Alternatively, the expectation could be replaced by using the outer measure μ_*^m of μ^m, replacing $\mathbf{E}(\Pi_H(\mathrm{x}))$ by

$$\mathbf{E}^*(\Pi_H(\mathrm{x})) = \sum_{k=1}^{2^m} k\, \mu_*^m \{\mathrm{x} \in X^m : \Pi_H(\mathrm{x}) = k\}.$$

Similar remarks apply to Π_{B_ϵ}.

The proof of Theorem 3.3 is quite involved, and we require some preliminary results to convert the proof to a simple counting argument.

Group action

A key technique in the proof is to use a group action on the product space to convert the problem to a combinatorial one.

The symmetric group of degree n, S_n, has a natural action on X^n. For any $\sigma \in S_n$ and $\mathrm{x} = (x_1, \ldots, x_n) \in X^n$, we define $\sigma\mathrm{x}$ by

$$\sigma\mathrm{x} = \left(x_{\sigma(1)}, \ldots, x_{\sigma(n)}\right).$$

That is, the entries of the vector are permuted according to σ. We make the following definitions.

Definition 3.4 *Let Λ be a subset of the full symmetric group of degree n, acting on X^n as above. For any subset A of X^n, and for any $\mathrm{x} \in X^n$, define*

$$\Omega_A(\Lambda, \mathrm{x}) = |\{\sigma \in \Lambda : \sigma\mathrm{x} \in A\}|$$

and

$$\Omega_A(\Lambda) = \sup\{\Omega_A(\Lambda, \mathrm{x}) : \mathrm{x} \in X^n\}.$$

□

The following lemma will prove useful. Given a set Λ of permutations of degree n, it enables us to express the measure of a μ^n-measurable subset A of X^n in terms of $\Omega_A(\Lambda, \mathrm{x})$ and to bound the measure of A in terms of the combinatorial parameter $\Omega_A(\Lambda)$.

Lemma 3.5 *Suppose that A is a μ^n-measurable subset of X^n and that Λ is any subset of the full symmetric group of degree n, acting on X^n in the natural way. Then $\Omega_A(\Lambda, \mathrm{x})$ is a measurable function, and*

$$\mu^n(A) = \frac{1}{|\Lambda|}\mathbf{E}(\Omega_A(\Lambda, \mathrm{x})) \leq \frac{\Omega_A(\Lambda)}{|\Lambda|},$$

where $\mathbf{E}(.)$ denotes expected value over X^n with respect to the product measure μ^n.

Proof For a subset S of X^n, let I_S denote the characteristic (or indicator) function of S. That is, I_S is the $\{0,1\}$-valued function on X^n such that $I_S(\mathrm{x}) = 1$ if and only if x belongs to S. If S is a measurable subset of X^n, then I_S is clearly a measurable function. Now, the symmetric group of degree n acts as a measure-preserving group of transformations of X^n, with respect to the product measure μ^n. That is, for any $\tau \in S_n$ and for any μ^n-measurable subset S of X^n, the set $\tau S = \{\tau\mathrm{x} : \mathrm{x} \in S\}$ is measurable and $\mu^n(\tau S) = \mu^n(S)$. It follows that

$$\Omega_A(\Lambda, \mathrm{x}) = \sum_{\sigma \in \Lambda} I_A(\sigma\mathrm{x}) = \sum_{\sigma \in \Lambda} I_{\sigma^{-1}A}(\mathrm{x}),$$

as a sum of measurable functions, is therefore a measurable function. Now,

$$\begin{aligned}
|\Lambda|\,\mu^n(A) &= \sum_{\sigma\in\Lambda} \mu^n(A) \\
&= \sum_{\sigma\in\Lambda} \mu^n\left(\sigma^{-1}A\right) \\
&= \sum_{\sigma\in\Lambda} \int_{X^n} I_{\sigma^{-1}A}(\mathrm{x})\, d\mu^n(\mathrm{x}) \\
&= \sum_{\sigma\in\Lambda} \int_{X^n} I_A(\sigma\mathrm{x})\, d\mu^n(\mathrm{x}) \\
&= \int_{X^n} \left(\sum_{\sigma\in\Lambda} I_A(\sigma\mathrm{x})\right) d\mu^n(\mathrm{x}) \\
&= \int_{X^n} \Omega_A(\Lambda,\mathrm{x})\, d\mu^n(\mathrm{x}) \\
&= \mathbf{E}\left(\Omega_A(\Lambda,\mathrm{x})\right),
\end{aligned}$$

and clearly

$$\mathbf{E}\left(\Omega_A(\Lambda,\mathrm{x})\right) \leq \Omega_A(\Lambda).$$

The result follows. □

Proof of the bounding theorem

Following [11], Theorem 3.3 is proved in two main stages. The first relates the measure of Q to the measure of J, and the second uses group action to bound the measure of J by means of a combinatorial argument.

Proposition 3.6 *With Q and J as before, for any positive integer $k > 1/\epsilon$ and for any r such that*

$$0 < r < 1 - \frac{1}{\sqrt{\epsilon k}},$$

the following holds:

$$\mu^m(Q) < \frac{\epsilon k(1-r)^2}{\epsilon k(1-r)^2 - 1}\,\mu^{m+k}(J) = C(r,k)\,\mu^{m+k}(J).$$

Proof The proof uses Chebyshev's inequality [14], which states that if $\eta > 0$ and Y is a bounded random variable with expectation zero then

$$\mathrm{Prob}\left(|Y| > \eta\right) \leq \frac{\sigma^2}{\eta^2},$$

where σ^2 is the variance of Y. For a particular $h \in B_\epsilon$, let

$$\epsilon_h = \mathrm{er}_\mu(h) > \epsilon.$$

Then,

$$\begin{aligned}\mu^k\left\{\mathrm{y} \in X^k : \mathrm{er}_\mathrm{y}(h) \le r\epsilon\right\} &= \mu^k\left\{\mathrm{y} \in X^k : \epsilon_h - \mathrm{er}_\mathrm{y}(h) \ge \epsilon_h - r\epsilon\right\} \\ &\le \mu^k\left\{\mathrm{y} \in X^k : |k\,\mathrm{er}_\mathrm{y}(h) - k\epsilon_h| \ge (\epsilon_h - r\epsilon)k\right\}.\end{aligned}$$

Now, $k\,\mathrm{er}_\mathrm{y}(h)$, the number of entries of y on which h and c disagree, is a binomially distributed random variable on X^k with expected value $\epsilon_h k$ and variance $\epsilon_h(1-\epsilon_h)k$. It follows, by Chebyshev's inequality, that this measure is at most

$$\begin{aligned}\frac{\epsilon_h(1-\epsilon_h)k}{((\epsilon_h - r\epsilon)k)^2} &\le \frac{\epsilon_h(1-\epsilon_h)}{(\epsilon_h - r\epsilon_h)(\epsilon - r\epsilon)k} \\ &= \frac{1-\epsilon_h}{\epsilon k(1-r)^2} \\ &\le \frac{1}{\epsilon k(1-r)^2}.\end{aligned}$$

Therefore, for any $h \in B_\epsilon$,

$$\mu^k\left\{\mathrm{y} \in X^k : \mathrm{er}_\mathrm{y}(h) > r\epsilon\right\} \ge 1 - \frac{1}{\epsilon k(1-r)^2} = C(r,k)^{-1}.$$

It follows that for any $\mathrm{x} \in Q$,

$$\begin{aligned}\int_{\mathrm{y}\in X^k} I_J(\mathrm{x},\mathrm{y})\,d\mu^k &= \mu^k\left\{\mathrm{y} \in X^k : \exists h \in B_\epsilon \cap H[\mathrm{x}] \text{ s.t. } \mathrm{er}_\mathrm{y}(h) > r\epsilon\right\} \\ &\ge \sup(\mu^k\left\{\mathrm{y} \in X^k : \mathrm{er}_\mathrm{y}(h) > r\epsilon\right\}) \\ &\ge C(r,k)^{-1},\end{aligned}$$

where the supremum is taken over all h in $B_\epsilon \cap H[\mathrm{x}]$.

By Fubini's theorem,

$$\begin{aligned}\mu^{m+k}(J) &= \int_{X^{m+k}} I_J(x_1,\ldots,x_{m+k})\,d\mu^{m+k} \\ &= \int_{\mathrm{x}\in X^m}\left(\int_{\mathrm{y}\in X^k} I_J(\mathrm{x},\mathrm{y})\,d\mu^k\right)d\mu^m \\ &\ge \int_{\mathrm{x}\in Q}\left(\int_{\mathrm{y}\in X^k} I_J(\mathrm{x},\mathrm{y})\,d\mu^k\right)d\mu^m.\end{aligned}$$

Therefore,

$$\mu^{m+k}(J) \ge \frac{1}{C(r,k)}\int_Q d\mu^m(\mathrm{x}) = \frac{1}{C(r,k)}\,\mu^m(Q),$$

from which the result follows. □

Proposition 3.7 *With J defined as above, and for r and k as in Proposition 3.6, for all $\mathrm{z} \in X^{m+k}$, we have*

$$\frac{\Omega_J(S_{m+k}, \mathrm{z})}{(m+k)!} \leq \Pi_{B_\epsilon}(\mathrm{z}) \binom{k}{\lceil r\epsilon k \rceil} \binom{m+k}{\lceil r\epsilon k \rceil}^{-1}.$$

Proof Fix $\mathrm{z} \in X^{m+k}$ and let

$$\{h_1, h_2, \ldots, h_t\} \subseteq B_\epsilon$$

be a complete set of representatives of B_ϵ for z. By this, we mean

(1) $$t = \Pi_{B_\epsilon}(\mathrm{z}),$$

(2) $$i \neq j \Longrightarrow h_i(\mathrm{z}) \neq h_j(\mathrm{z}),$$

(3) $$\{h_i(\mathrm{z}) : 1 \leq i \leq t\} = \{h(\mathrm{z}) : h \in B_\epsilon\}.$$

For each i between 1 and t, define

$$J^i = \left\{\mathrm{xy} \in X^{m+k} : h_i \in H[\mathrm{x}],\ \mathrm{er}_{\mathrm{y}}(h_i) > r\epsilon\right\},$$

and, for $\mathrm{z} \in X^{m+k}$, let

$$\Omega_J^i(\mathrm{z}) = \Omega_{J^i}(S_{m+k}, \mathrm{z}) = \left|\left\{\sigma \in S_{m+k} : \sigma\mathrm{z} \in J^i\right\}\right|.$$

Then

$$\begin{aligned}
\Omega_J(S_{m+k}, \mathrm{z}) &= \left|\{\sigma \in S_{m+k} : \sigma\mathrm{z} \in J\}\right| \\
&= \left|\left\{\sigma \in S_{m+k} : \sigma\mathrm{z} \in \bigcup_{i=1}^{t} J^i\right\}\right| \\
&\leq \sum_{i=1}^{t} \left|\left\{\sigma \in S_{m+k} : \sigma\mathrm{z} \in J^i\right\}\right| \\
&= \sum_{i=1}^{t} \Omega_J^i(\mathrm{z}).
\end{aligned}$$

Suppose that $\Omega_J^i(z) \neq 0$. Then there is $\sigma \in S_{m+k}$ such that $\sigma z \notin J^i$. Let $l = k\,\mathrm{er}_z(h_i)$ be the number of entries of z on which h_i and c disagree. This is an integer and $\mathrm{er}_z(h) > r\epsilon$; thus, $l \geq \lceil r\epsilon k \rceil$. If $\tau \in S_{m+k}$ is such that $\tau z \in J^i$, then τ must permute the entries of the vector z in such a way that the l entries on which h_i and c disagree are among the last k entries of the vector τz. The number of permutations τ for which this is the case is

$$\binom{k}{l} l!\,(m+k-l)!.$$

Therefore,

$$\begin{aligned}\frac{\Omega_J^i(z)}{(m+k)!} &= \frac{k(k-1)\cdots(k-l+1)}{(m+k)(m+k-1)\cdots(m+k-l+1)} \\ &= \left(\frac{k}{m+k}\right)\left(\frac{k-1}{m+k-1}\right)\cdots\left(\frac{k-l+1}{m+k-l+1}\right).\end{aligned}$$

Each term in this product is less than one and $l \geq \lceil r\epsilon k \rceil$, so an upper bound for the product is obtained by putting $l = \lceil r\epsilon k \rceil$ in the right-hand side, giving

$$\frac{\Omega_J^i(z)}{(m+k)!} \leq \binom{k}{\lceil r\epsilon k \rceil}\binom{m+k}{\lceil r\epsilon k \rceil}^{-1}.$$

Therefore,

$$\frac{\Omega_J(S_{m+k}, z)}{(m+k)!} \leq \sum_{i=1}^{t} \frac{\Omega_J^i(z)}{(m+k)!} \leq \Pi_{B_\epsilon}(z)\binom{k}{\lceil r\epsilon k \rceil}\binom{m+k}{\lceil r\epsilon k \rceil}^{-1},$$

and the Proposition is proved. □

We are now in a position to prove Theorem 3.3.

Proof of Theorem 3.3 By Lemma 3.5 and Proposition 3.6, taking $n = m+k$ and $\Lambda = S_{m+k}$,

$$\mu^m(Q) \leq C(r,k)\,\mu^{m+k}(J) = C(r,k)\frac{1}{(m+k)!}\mathbf{E}\left(\Omega_J(S_{m+k}, z)\right).$$

Now, by Proposition 3.7,

$$\frac{\Omega_J(S_{m+k}, z)}{(m+k)!} \leq \Pi_{B_\epsilon}(z)\binom{k}{\lceil r\epsilon k \rceil}\binom{m+k}{\lceil r\epsilon k \rceil}^{-1}.$$

If Π_{B_ϵ} is a measurable function then its expected value is defined, and the first part of the theorem follows if in addition we use the obvious fact that

$$\mathbf{E}(\Pi_{B_\epsilon}(\mathrm{x})) \leq \mathbf{E}(\Pi_H(\mathrm{x}))\,,$$

assuming that Π_H is measurable. The second part of the theorem follows on using Lemma 3.5 to give

$$\begin{aligned}
\mu^{m+k}(J) &\leq \sup\left\{\frac{\Omega_J(S_{m+k},\mathrm{z})}{(m+k)!} : \mathrm{z}\in X^{m+k}\right\} \\
&\leq \sup\left\{\Pi_{B_\epsilon}(\mathrm{z})\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1} : \mathrm{z}\in X^{m+k}\right\} \\
&= \Pi_{B_\epsilon}(m+k)\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1}
\end{aligned}$$

which, of course, is at most

$$\Pi_H(m+k)\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1}.$$

□

3.3 Learnability in Spaces of Finite VC Dimension

A distribution-independent bound

Theorem 3.3 can be used for obtaining distribution-independent learnability results. However, it is the fourth, and weakest, assertion that must be used since both the set B_ϵ and the expectations of the index functions depend on the probability measure μ. Applying the theorem, we have

Corollary 3.8 *With Q, r, k as before,*

$$\mu^m(Q) < C(r,k)\,\Pi_H(m+k)\left(\frac{k}{m+k}\right)^{r\epsilon k}.$$

Proof We use the fact, from Theorem 3.3, that

$$\mu^m(Q) \leq C(r,k)\,\Pi_H(\mathrm{x})\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1}.$$

Now,

$$\begin{aligned}\binom{k}{\lceil r\epsilon k\rceil}\binom{m+k}{\lceil r\epsilon k\rceil}^{-1} &= \frac{k(k-1)\ldots(k-\lceil r\epsilon k\rceil+1)}{(m+k)(m+k-1)\ldots(m+k-\lceil r\epsilon k\rceil+1)} \\ &= \left(\frac{k}{m+k}\right)\left(\frac{k-1}{m+k-1}\right)\cdots\left(\frac{k-\lceil r\epsilon k\rceil+1}{m+k-\lceil r\epsilon k\rceil+1}\right) \\ &< \left(\frac{k}{m+k}\right)^{\lceil r\epsilon k\rceil} \\ &\leq \left(\frac{k}{m+k}\right)^{r\epsilon k},\end{aligned}$$

and the result follows. □

Suppose that $m \geq 8/\epsilon$, and $k = m$. Then

$$\frac{1}{2} < 1 - \frac{1}{\sqrt{\epsilon k}},$$

so we may take $k = m$ and $r = 1/2$ in Corollary 3.8. This gives:

Corollary 3.9 *For $m \geq 8/\epsilon$,*

$$\mu^m(Q) < 2\,\Pi_H(2m)\,2^{-\epsilon m/2}.$$

Proof We have

$$C\left(\frac{1}{2}, m\right) = \frac{\epsilon m\left(\frac{1}{2}\right)^2}{\epsilon m\left(\frac{1}{2}\right)^2 - 1} \leq 2,$$

and

$$\left(\frac{k}{m+k}\right)^{r\epsilon k} = 2^{-\epsilon m/2}.$$

□

This is essentially the result in [16] and [11]. We shall show that other choices of r and k provide better bounds. The following result will be useful in our analysis:

Lemma 3.10 *For any ϵ, r with $0 < \epsilon, r < 1$, and for any positive integers m and k,*

$$\left(\frac{k}{m+k}\right)^{r\epsilon k} < \exp\left\{-r\epsilon\frac{km}{m+k}\right\}.$$

Proof For $0 < y < 1$, from the power series expansion of $\log(1-y)$, we have

$$\frac{1}{y}\log(1-y) < -1.$$

It follows that for all $x > 1$,

$$\left(1-\frac{1}{x}\right)^{x} < e^{-1}.$$

Therefore,

$$\begin{aligned}\left(\frac{k}{m+k}\right)^{r\epsilon k} &= \left(1-\frac{m}{m+k}\right)^{\frac{m+k}{m}r\epsilon\frac{km}{m+k}} \\ &< \exp\left\{-r\epsilon\frac{km}{m+k}\right\}.\end{aligned}$$

□

Suppose now that H has finite VC dimension $d \geq 2$. By Sauer's Lemma, we have

Proposition 3.11 *With Q, r, k as before, if H has finite VC dimension d,*

$$\mu^m(Q) < C(r,k)\left(\frac{e(m+k)}{d}\right)^{d}\exp\left\{-r\epsilon\frac{km}{m+k}\right\},$$

whenever $m + k > d$. □

Choosing r and k.

We aim to show that it is possible to choose values of r and k which give better bounds than previously obtained. Motivated by Proposition 3.11 and the fact that the real function

$$f(x) = \left(\frac{(m+x)}{d}\right)^d \exp\left\{-r\epsilon\frac{xm}{m+x}\right\}$$

is minimized when

$$x = m\left(\frac{r\epsilon m}{d} - 1\right),$$

we choose

$$k = \left\lceil m\left(\frac{r\epsilon m}{d} - 1\right)\right\rceil.$$

Before choosing a value of r, we require the following result:

Lemma 3.12 *For any $\beta \geq 2$ and $\epsilon m \geq 4d$ the equations*

$$x = m\left(\frac{r\epsilon m}{d} - 1\right),$$

$$r = 1 - \sqrt{\frac{\beta}{\epsilon x}},$$

have a solution (x, r) with $x \geq m$.

Proof The real number $x > 0$ is a solution to the equations if and only if

$$\frac{\epsilon m^2}{d} - \frac{\epsilon m^2}{d}\sqrt{\frac{\beta}{\epsilon x}} - m = x.$$

Let

$$g(y) = y^3 - \left(\frac{\epsilon m^2}{d} - m\right)y - \frac{\epsilon m^2}{d}\sqrt{\frac{\beta}{\epsilon}}.$$

Then the original equations have a solution for $x \geq m$ if and only if $g(y) = 0$ for some $y \geq \sqrt{m}$. But $\epsilon m \geq 4d$. Therefore,

$$g(\sqrt{m}) \leq m\sqrt{m} - 3m\sqrt{m} < 0.$$

Since $g(y)$ tends to infinity with y, it follows that g has a zero y such that $y \geq \sqrt{m}$. □

We obtain the following distribution-independent bound on the probability of presenting a bad sample.

Theorem 3.13 *Suppose that H is a hypothesis space over an input space X and that H has finite VC dimension $d \geq 2$. Let μ be any probability measure on X and let $m \geq 4d/\epsilon$. Then we have*

$$\mu^m \{x \in X^m : \mathrm{haz}_\mu(H[x]) > \epsilon\} < \left(\frac{d}{d-1}\right) \epsilon^d \left(\frac{e^2 m}{d}\right)^{2d} \exp(-\epsilon m).$$

Proof The required probability is the measure of Q. Let $x \geq m$ be a solution to the equations of Lemma 3.12, and let

$$k = \lceil x \rceil, \quad r = 1 - \sqrt{\frac{d}{\epsilon x}}.$$

Then

$$r \leq 1 - \sqrt{\frac{d}{\epsilon k}} < 1 - \frac{1}{\sqrt{\epsilon k}},$$

and

$$\mu^m(Q) < C(r,k) \left(\frac{e(m+x+1)}{d}\right)^d \exp\left\{-r\epsilon \frac{km}{m+k}\right\}.$$

It can easily be shown that

$$m + x + 1 = \frac{\epsilon r m^2}{d} + 1 < \frac{\epsilon m^2}{d}.$$

Further, since $k \geq x$,

$$C(r,k) \leq C(r,x) = \left(\frac{d}{d-1}\right).$$

Now,

$$r = 1 - \sqrt{\frac{d}{\epsilon x}} \geq 1 - \sqrt{\frac{d}{\epsilon m}} \geq 1 - \sqrt{\frac{d}{4d}} = \frac{1}{2}.$$

Hence

$$x = m\left(\frac{r\epsilon m}{d} - 1\right) \geq \frac{\epsilon m^2}{2d} - m \geq \frac{\epsilon m^2}{4d},$$

this last inequality because $\epsilon m \geq 4d$.

Now, $k \geq x$ and so

$$\begin{aligned}\frac{r\epsilon k}{m+k} &\geq \frac{r\epsilon x}{m+x} \\ &= \frac{xd}{m} \\ &= d\left(\frac{r\epsilon m}{d} - 1\right) \\ &= r\epsilon m - d \\ &= \epsilon m\left(1 - \sqrt{\frac{d}{\epsilon x}}\right) - d \\ &= \epsilon m - \epsilon m\sqrt{\frac{d}{\epsilon x}} - d.\end{aligned}$$

But $x \geq \epsilon m^2/4d$ and so

$$\exp\left\{-r\epsilon\frac{mk}{m+k}\right\} \leq e^d\, e^{2d} \exp(-\epsilon m).$$

It follows that

$$\mu^m(Q) < \left(\frac{d}{d-1}\right)\left(\frac{e^4\epsilon m^2}{d^2}\right)^d \exp(-\epsilon m).$$

□

A learnability theorem

We can use Theorem 3.13 to obtain learnability results and sample-size bounds for hypothesis spaces of finite VC dimension. In order to do so, we require one further lemma.

Lemma 3.14 *For any $\alpha > 0$ and for $x \geq 0$,*

$$\log x \leq \left(\log\left(\frac{1}{\alpha}\right) - 1\right) + \alpha x.$$

Proof By elementary calculus, the real function

$$f(x) = \log x - \alpha x$$

is maximized when $x = 1/\alpha$, and its maximum value is $-\log\alpha - 1$. □

Theorem 3.15 *Let H be a hypothesis space of finite VC dimension $d \geq 2$. Then H is learnable. Given an accuracy parameter $0 < \epsilon < 1$ and a confidence parameter $0 < \delta < 1$, a suitable sufficient sample-size for learnability to accuracy ϵ with confidence $1 - \delta$ is*

$$m_0(\epsilon, \delta) = \left\lceil \frac{1}{\epsilon(1 - \sqrt{\epsilon})} \left(2d \log \left(\frac{2e}{\epsilon} \right) + \log \left(\frac{d/(d-1)}{\delta} \right) \right) \right\rceil .$$

Proof From Theorem 3.13,

$$\mu^m(Q) < \left(\frac{d}{d-1} \right) \epsilon^d \left(\frac{e^2 m}{d} \right)^{2d} \exp(-\epsilon m),$$

for $m \geq 4d/\epsilon$. It follows that $\mu^m(Q)$ tends to zero as m tends to infinity, and that the rate of convergence is independent of the target concept and the probability measure on X. That is, H is learnable. More specifically, we show that if $0 < \epsilon, \delta < 1$ and

$$m \geq \frac{1}{\epsilon(1 - \sqrt{\epsilon})} \left(2d \log \left(\frac{2e}{\epsilon} \right) + \log \left(\frac{d/(d-1)}{\delta} \right) \right),$$

then

$$\left(\frac{d}{d-1} \right) \left(\frac{e^2 m}{d} \right)^{2d} \epsilon^d \exp(-\epsilon m) \leq \delta$$

and, consequently, by Theorem 3.13, $\mu^m(Q) < \delta$.

Let $D = d/(d-1)$. Then,

$$D \left(\frac{e^2 m}{d} \right)^{2d} \epsilon^d \exp(-\epsilon m) \leq \delta$$

$$\iff \log D + 4d + 2d \log m - 2d \log d + d \log \epsilon - \epsilon m \leq \log \delta$$

$$\iff \epsilon m \geq \log \left(\frac{D}{\delta} \right) + 4d - 2d \log d + d \log \epsilon + 2d \log m.$$

Let $\alpha = q\epsilon/2d$ in Lemma 3.14, with $0 < q < 1$. We have

$$\begin{aligned} 2d \log m &\leq 2d \left(\log \left(\frac{2d}{q\epsilon} \right) - 1 \right) + q\epsilon m \\ &= 2d \log(2d) - 2d \log q - 2d \log \epsilon - 2d + q\epsilon m. \end{aligned}$$

Therefore, it suffices to have

$$\begin{aligned}\epsilon m(1-q) &\geq \log\left(\frac{D}{\delta}\right) - d\log\epsilon - 2d\log q - 2d\log d + 2d\log 2d + 2d \\ &= \log\left(\frac{D}{\delta}\right) + d\log\left(\frac{1}{\epsilon}\right) - 2d\log q + 2d + 2d\log 2.\end{aligned}$$

Observing that $2d + 2d\log 2 = 2d\log(2e)$, a sufficient sample-size is

$$\left\lceil \frac{1}{\epsilon(1-q)}\left(d\log\left(\frac{4e^2}{\epsilon q^2}\right) + \log\left(\frac{D}{\delta}\right)\right)\right\rceil.$$

Choosing $q = \sqrt{\epsilon}$ yields the result. □

The sample-size bound given in Theorem 3.15 improves upon the bound

$$m_0(\epsilon,\delta) = \left\lceil \frac{4}{\epsilon}\left(2d\log_2\left(\frac{13}{\epsilon}\right) + \log_2\left(\frac{2}{\delta}\right)\right)\right\rceil,$$

in [16].

3.4 Lower Bounds on Necessary Sample-Size

We have seen that finite VC dimension is a sufficient condition on a space H for H to be learnable. There is a strong converse to this; if a hypothesis space H is learnable then it must have finite VC dimension. This result can be proved quite easily (see Chapter 6), but it also follows from lower bounds, involving the VC dimension, on necessary sample-size. We prove one easy such bound and thereby show that finite VC dimension of the hypothesis space is necessary for learnability. We then describe lower bounds of Blumer *et al* [11] and Ehrenfeucht *et al* [13].

The following lower bound result is easily obtained.

Theorem 3.16 *Suppose that H is a hypothesis space over an input space X and that H has finite VC dimension d. Let $0 < \epsilon < 1$. Then, there is a probability measure μ on X such that for any target concept $c \in H$ and for any positive integer m with*

$$m < d(1-\epsilon),$$

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} = 1.$$

Proof Let $c \in H$ be any target concept and suppose that m is a positive integer less than $d(1-\epsilon)$. Since H has VC dimension d, there is a set S of d points of X shattered by H. Let the probability measure μ be uniform on S and zero elsewhere. That is, define μ by defining, for a measurable subset A of X,

$$\mu(A) = \frac{1}{d}\left|A \cap S\right|.$$

Let

$$\mathrm{x} = (x_1, x_2, \ldots, x_m) \in S^m.$$

Then, since H shatters S, there is $h \in H$ such that h agrees with c on $x_1, x_2, \ldots, x_m$, but disagrees with c on each of the other $d-m$ points of S. Therefore, there is $h \in H[\mathrm{x}]$ such that

$$\mathrm{er}_\mu(h) = (d-m)\frac{1}{d} > \frac{d-(1-\epsilon)d}{d} = \epsilon,$$

and the result follows on observing that

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} \geq \mu^m(S^m) = 1.$$

□

This result shows that if we aim to learn H to accuracy ϵ with *any* degree of confidence, we need a sample-size greater than $d(1-\epsilon)$.

We have

Corollary 3.17 *If the hypothesis space H has infinite VC dimension then H is not learnable.*

Proof Suppose that H has infinite VC dimension and let $c \in H$ be any target concept. For each positive integer d there is a set X_d of d points shattered by H. The above result shows that if μ_d is the probability measure on X which is uniform on X_d and zero outside X_d, then for all $m < d/2$,

$$\mu_d^m \left\{ \mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \frac{1}{2} \right\} = 1.$$

Suppose that H is learnable. Then, with the function $m_0(\epsilon, \delta)$ as in the definition of learnability, we must have

$$m_0\left(\frac{1}{2}, \frac{1}{2}\right) \geq \frac{d}{2},$$

and this must hold for any positive integer d. This is a clear impossibility, and the result follows. □

Ehrenfeucht *et al* [13] have (essentially) given the following stronger lower bound, which we shall not prove here.

Theorem 3.18 *Suppose that H is a hypothesis space over an input space X and that H has finite VC dimension d. Suppose that $0 < \epsilon \leq 1/8$. Then there is a probability measure μ on X such that for any positive integer m with*

$$m < \frac{d-1}{32\epsilon},$$

$$\mu^m \left\{ \mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon \right\} > \frac{1}{100}.$$

□

Neither of these bounds has any explicit dependence on δ. The following result of Blumer *et al* [11] gives a sample-size lower bound that depends on ϵ and δ, but not on the VC dimension of the hypothesis space.

Theorem 3.19 *Suppose H is a hypothesis space over an input space X and that $0 < \epsilon, \delta < 1$. Then there is a probability distribution μ on X such that for any positive integer m with*

$$m < \frac{(1-\epsilon)}{\epsilon} \log\left(\frac{1}{\delta}\right),$$

we have

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} > \delta.$$

Proof Suppose firstly that there are two hypotheses h and g in H such that for some ~~x, y~~ $a, b \in X$, $h(a) = g(a) = 1$ and $g(b) = 1$, $h(b) = 0$ (that is, h and g are neither equal nor disjoint). Let the measure μ be such that $\mu(\{b\}) = \epsilon$, $\mu(\{a\}) = 1 - \epsilon$, and μ is zero elsewhere on X. Let the target concept be h. Suppose an m-sample x is drawn randomly, according to μ. The probability that each entry of x is a is $(1-\epsilon)^m$, and if

$$m < \frac{1}{-\log(1-\epsilon)} \log\left(\frac{1}{\delta}\right),$$

this is at least δ. In this case, g is consistent with h on the sample, but has error ϵ (the probability of b). Now,

$$\frac{1}{-\log(1-\epsilon)} > \frac{1-\epsilon}{\epsilon},$$

and the result follows for this case.

The only remaining case to consider (since H is non-trivial) is when there are distinct $h, g \in H$ such that $h(x) = 1$ implies $g(x) = 0$, $g(x) = 1$ implies $h(x) = 0$ (that is, h, g are disjoint), and there is $a \in X$ such that $h(a) = g(a) = 0$ (that is, h and g are not complementary). Now, h and g cannot both be the identically zero function on X (for, they are not equal) and so, without loss of generality, we may assume that there is some $b \in X$ such that $h(b) = 1$. Define the probability measure μ as before. The same analysis now applies. □

The main sample-size bounds presented in this chapter can be summarized in the following.

Theorem 3.20 *Let H be a hypothesis space over input space X. Then H is learnable if and only if H has finite VC dimension. If H has finite VC dimension $d \geq 2$ then, given $0 < \epsilon, \delta < 1$, there is $m_0 = m_0(\epsilon, \delta)$ such that*

$$m > m_0 \Longrightarrow \mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu (H[\mathrm{x}]) > \epsilon\} < \delta.$$

The sample-size $m_0(\epsilon, \delta)$ satisfies

$$m_0(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon(1-\sqrt{\epsilon})} \left(2d \log \left(\frac{2e}{\epsilon} \right) + \log \left(\frac{d/(d-1)}{\delta} \right) \right) \right\rceil$$

and, for $\delta \leq 1/100$,

$$m_0(\epsilon, \delta) > \max \left(\frac{(1-\epsilon)}{\epsilon} \log \left(\frac{1}{\delta} \right), \frac{d-1}{32\epsilon} \right).$$

□

Chapter 4
Relative Frequencies and Probabilities

4.1 Introduction

Definitions

Throughout this chapter, (S, Σ, ν) will be a probability space and $\mathcal{C}$ will be a collection of sets from the σ-algebra Σ. Thus, in this probabilistic setting, S may be thought of as a set of elementary events and $\mathcal{C}$ as a collection of random events. We shall need the theory only for countable S, in which case Σ will consist of all subsets of S, and for S a subset of some real Euclidean space, in which case Σ will be the induced Borel σ-algebra. The class $\mathcal{C}$ must satisfy certain measurability conditions, which we shall not include here. For details of these "permissibility" conditions on $\mathcal{C}$, see [27]. We shall assume that all classes under discussion here are permissible and that all sets we require to measure are therefore measurable.

A *sample* from S of length m is a vector $\mathrm{y} = (y_1, y_2, \ldots, y_m) \in S^m$. The *relative frequency* of occurence of event $A \in \mathcal{C}$ on y is defined to be

$$\mathbf{P}_{\mathbf{y}}(A) = \frac{1}{m} |\{i : y_i \in A\}| .$$

This is the empirical estimate on sample y of the probability of A. Further, we let $I(\mathrm{y}) = \{y_i : 1 \leq i \leq m\}$ be the set of entries of y.

Uniform convergence of relative frequencies

We say that the relative frequency of an event A in a class $\mathcal{C}$ of events tends (in probability) to the probability of A as the sample-size tends to infinity if for any $\eta > 0$,

$$\nu^m \{\mathrm{x} \in S^m : |\mathrm{P}_\mathrm{x}(A) - \nu(A)| > \eta\} \to 0 \text{ as } m \to \infty.$$

Classical theorems of probability theory, such as Hoeffding's inequality, assure us of the convergence (in probability) of the relative frequency of an event to the probability of the event.

Whilst the relative frequencies of the events in $\mathcal{C}$ converge to their probabilities, there may be no bound uniform over $\mathcal{C}$ on the rate of this convergence. We say that the relative frequencies of events $A \in \mathcal{C}$ *converge uniformly over* $\mathcal{C}$ (in probability) to their probabilities if for any $\eta > 0$

$$\nu^m \left\{\mathrm{x} \in S^m : \sup_{\mathcal{C}} |\mathrm{P}_\mathrm{x}(A) - \nu(A)| > \eta\right\} \to 0 \text{ as } m \to \infty.$$

Thus, the relative frequencies converge to the probabilities uniformly over $\mathcal{C}$ if and only if the rate of convergence for each event can be bounded by a quantity depending only on the class $\mathcal{C}$. Clearly, if there are only a finite number of events in $\mathcal{C}$, then we have such uniform convergence. However, when $\mathcal{C}$ is infinite a more sophisticated theory is necessary. The VC dimension of the class $\mathcal{C}$ is of great importance here.

Our aim in this chapter is to give simple new proofs of two theorems, due to Vapnik [35], which provide bounds on the probability of a given deviation of relative frequencies from probabilities. These theorems prove uniform convergence of relative frequencies to probabilities over classes $\mathcal{C}$ of finite VC dimension and the theory of the preceeding chapter (with slightly weaker bounds) follows immediately from the second of these theorems. The results show not only that if the class $\mathcal{C}$ has finite VC dimension then the relative frequencies of events in $\mathcal{C}$ converge uniformly over $\mathcal{C}$ to their probabilities, but also that the

rate of convergence can be bounded independently of the probability measure ν. The techniques we use are similar to those used in the preceeding chapter.

Two bounding theorems

The bounds we derive are due to Vladimir Vapnik [35] (although our results have slightly different constants). The first is a bound on the probability that the relative frequency of an event A in $\mathcal{C}$ differs from the probability of A by more than a certain amount.

Theorem 4.1 [Bound One] *With the above definitions, for any $\eta > 0$ and for any positive integer m,*

$$\nu^m \left\{ x \in S^m : \sup_{\mathcal{C}} |\mathbf{P}_x(A) - \nu(A)| > \eta \right\} \leq 4\, \Pi_{\mathcal{C}}(2m) \exp\left(-\frac{1}{8}\eta^2 m\right).$$

□

The second result concerns relative deviation rather than absolute deviation.

Theorem 4.2 [Bound Two] *With the above definitions, for any $\eta > 0$ and for any positive integer m,*

$$\nu^m \left\{ x \in S^m : \sup_{\mathcal{C}} \frac{\nu(A) - \mathbf{P}_x(A)}{\sqrt{\nu(A)}} > \eta \right\} \leq 4\, \Pi_{\mathcal{C}}(2m) \exp\left(-\frac{1}{4}\eta^2 m\right).$$

□

We shall prove each of these results in two different ways. Although Bound One follows from Bound Two (indeed, the obvious stronger result follows), we include a proof of Bound One for completeness and to illustrate the common proof techniques.

When $\mathcal{C}$ has finite VC dimension it follows by Sauer's Lemma that the growth function is polynomially bounded and both these bounds tend to zero as m tends to infinity. That is,

Corollary 4.3 *If $\mathcal{C}$ has finite VC dimension, then the relative frequencies of events in $\mathcal{C}$ converge uniformly over $\mathcal{C}$ to their probabilities.* □

Clearly, also, the bounds are independent of the measure ν. Thus these bounds are precisely the types of results we need for learning applications. These applications are discussed later, where it will become apparent that Bounds One and Two both imply learnability results, but that the consequent upper bound on sufficient sample-size implied by Bound Two is significantly less than that implied by Bound One.

4.2 Proof Techniques

In this section, we describe how the results are proved. We leave the technicalities to the next section; the aim here is to give an idea of the techniques involved.

Symmetrization

As in [35], [27], [18] and the preceeding chapter, the desired probability is first bounded in terms of the probability of an event in some higher-dimensional product space, this event being "empirically" based on two samples. We shall follow Pollard, and call this technique *symmetrization.* In what follows, we shall often write a vector in S^{2m} in the form xy, where $\mathrm{x}, \mathrm{y} \in S^m$, and we assume (by the permissibility of H) that all sets discussed are measurable. The symmetrization results are as follows.

Proposition 4.4 *With the above notation, for $\eta > 0$, let*

$$Q = \left\{ \mathrm{x} \in S^m : \sup_{C} |\mathbf{P}_{\mathrm{x}}(A) - \nu(A)| > \eta \right\} \subseteq S^m,$$

and

$$R = \left\{ \mathrm{xy} \in S^{2m} : \sup_{C} |\mathbf{P}_{\mathrm{x}}(A) - \mathbf{P}_{\mathrm{y}}(A)| > \frac{\eta}{2} \right\} \subseteq S^{2m}.$$

Then, for $m \geq 2/\eta^2$,

$$\nu^m(Q) \leq 2\,\nu^{2m}(R).$$

Proposition 4.5 *With the above notation, for $\eta > 0$, let*

$$V = \left\{ \mathrm{x} \in S^m : \sup_c \frac{\nu(A) - \mathbf{P}_{\mathrm{x}}(A)}{\sqrt{\nu(A)}} > \eta \right\} \subseteq S^m,$$

and

$$W = \left\{ \mathrm{xy} \in S^{2m} : \sup_c (\mathbf{P}_{\mathrm{y}}(A) - \mathbf{P}_{\mathrm{x}}(A)) > \eta \sqrt{\mathbf{P}_{\mathrm{xy}}(A)} \right\} \subseteq S^{2m}.$$

Then, for $m \geq 2/\eta^2$,

$$\nu^m(V) \leq 4\, \nu^{2m}(W).$$

□

The swapping subgroup and combinatorial bounding

As in the previous chapter, after symmetrization, we prove the results by using combinatorial arguments arising from consideration of a group action. We consider the natural action of a group of permutations of degree $2m$ on the vectors of S^{2m}. The particular group we shall use is the "swapping" subgroup of the full symmetric group of degree $2m$. The swapping subgroup was introduced in this context by Pollard [27] and greatly simplifies the counting arguments required.

Definition 4.6 *The swapping subgroup, $\Lambda = \Lambda_{2m}$, is that subgroup of the full symmetric group of degree $2m$ which is generated by the transpositions $(j, m+j)$, for j between 1 and m. That is,*

$$\Lambda = \langle (j, m+j) : 1 \leq j \leq m \rangle \leq S_{2m}.$$

□

Note that $|\Lambda| = 2^m$. Having symmetrized, we use the group action to bound the measures of R and W, using Lemma 3.5.

We let $\Omega_R(\Lambda, z)$ be the number of permutations $\sigma \in \Lambda$ such that $\sigma z \in R$. Lemma 3.5 then shows that

$$\nu^{2m}(R) \leq \frac{\Omega_R(\Lambda)}{|\Lambda|},$$

where

$$\Omega_R(\Lambda) = \sup\left\{\Omega_R(\Lambda, z) : z \in S^{2m}\right\}.$$

The next step is to fix, arbitrarily, $z = (z_1, ..., z_{2m}) \in S^{2m}$, and bound $\Omega_R(\Lambda, z)$ independently of z. This yields an upper bound for $\Omega_R(\Lambda)$, and hence for $\nu^{2m}(R)$ and $\nu^m(Q)$. We treat V and W in the same way.

Let $\{A_1, ..., A_t\}$ be a complete set of distinct representatives of $\mathcal{C}$ for z. That is,

(1) $$t = \Pi_{\mathcal{C}}(z),$$

(2) $$i \neq j \Longrightarrow A_i \cap I(z) \neq A_j \cap I(z),$$

(3) $$\{A_i \cap I(z) : 1 \leq i \leq 2m\} = \{A \cap I(z) : A \in \mathcal{C}\}.$$

Thus, the sets $A_i \cap I(z)$, $(1 \leq i \leq t)$ form a complete repetition-free list of all sets of the form $A \cap I(z)$ with A in $\mathcal{C}$.

For each i between 1 and t, we define the sets R^i and W^i to be the events R and W restricted to A_i. By this we mean

$$R^i = \left\{xy \in S^{2m} : |\mathbf{P}_x(A_i) - \mathbf{P}_y(A_i)| > \eta\right\},$$

and

$$W^i = \left\{xy \in S^{2m} : (\mathbf{P}_y(A_i) - \mathbf{P}_x(A_i)) > \eta\sqrt{\mathbf{P}_{xy}(A_i)}\right\}.$$

As in Definition 3.4, we let

$$\Omega_R^i(z) = \Omega_{R^i}(\Lambda, z) = \left|\{\sigma \in \Lambda : \sigma z \in R^i\}\right|.$$

Then, by an argument similar to that given in Chapter 3,

$$\begin{aligned}\Omega_R(\Lambda, z) &= |\{\sigma \in \Lambda : \sigma z \in R\}| \\ &= \left|\left\{\sigma \in \Lambda : \sigma z \in \bigcup_{i=1}^{t} R^i\right\}\right| \\ &\leq \sum_{i=1}^{t} |\{\sigma \in \Lambda : \sigma z \in R^i\}| \\ &= \sum_{i=1}^{t} \Omega_R^i(z).\end{aligned}$$

Thus, we can bound the quantity

$$\frac{\Omega_R(\Lambda, z)}{|\Lambda|},$$

and hence $\nu^{2m}(R)$, by bounding $\Omega_R^i(z)$ for each relevant i. Specifically, if

$$\frac{\Omega_R^i(z)}{|\Lambda|} \leq B(m)$$

for all $z \in S^{2m}$, then

$$\frac{\Omega_R(\Lambda, z)}{|\Lambda|} \leq tB(m) = \Pi_{\mathcal{C}}(z)B(m) \leq \Pi_{\mathcal{C}}(2m)B(m).$$

The same analysis applies to W. We call bounding

$$\frac{\Omega_R(\Lambda, z)}{|\Lambda|}, \quad \frac{\Omega_W(\Lambda, z)}{|\Lambda|}$$

in this manner *combinatorial bounding.*

The combinatorial bounding results are as follows.

Theorem 4.7 *With the above notation, for any $z \in S^{2m}$,*

$$\frac{\Omega_R(\Lambda, z)}{|\Lambda|} \leq 2\,\Pi_{\mathcal{C}}(2m) \exp\left(-\frac{1}{8}\eta^2 m\right).$$

□

Theorem 4.8 *With the above notation, for any $z \in S^{2m}$,*

$$\frac{\Omega_W(\Lambda, z)}{|\Lambda|} \leq \Pi_{\mathcal{C}}(2m) \exp\left(-\frac{1}{4}\eta^2 m\right).$$

Bounds One and Two follow from the symmetrization and the combinatorial bounding results. Indeed, combining these results implies that Bounds One and Two hold for any $m \geq 2/\eta^2$, while the bounds hold trivially for m less than $2/\eta^2$, since in this case the right-hand side of each of the bounds is greater than one.

4.3 Proofs

In this section, we prove the symmetrization and combinatorial bounding results described in the last section. Our symmetrization proofs are essentially those given by Vapnik, but the combinatorial bounding proofs are far simpler than those he gives, using the action of the swapping subgroup rather than the full symmetric group.

Exponential inequalities

For each of the bounds, we perform the combinatorial bounding in two ways. For this, we require two inequalities. The first, which appears in [23], is a bound on the tail of the binomial series and the second is Hoeffding's inequality, as stated in [27]. The first can be regarded as a special case of Hoeffding's inequality, but we derive it from a better known result of Chernoff [12].

Proposition 4.9 *For any positive integer n, and for any $\lambda < n/2$,*

$$\sum_{i<\frac{1}{2}n-\lambda} \binom{n}{i} < 2^n \exp\left(-2\lambda^2/n\right)$$

Proof Denote the sum by $\sum$. We use a bound of Chernoff [12]: For any $0 < p < 1$ and any positive integer n,

$$\sum_{i\leq k} \binom{n}{i} p^i (1-p)^{n-i} \leq \exp\left\{(n-k)\log\left(\frac{n(1-p)}{(n-k)}\right) + k\log\left(\frac{np}{k}\right)\right\}.$$

Putting $p = 1/2$, and $k = \frac{1}{2}n - \lambda$, this gives

$$2^{-n} \sum \leq \exp\left\{\left(\frac{1}{2}n + \lambda\right)\log\left(\frac{n}{n+2\lambda}\right) + \left(\frac{1}{2}n - \lambda\right)\log\left(\frac{n}{n-2\lambda}\right)\right\}$$
$$= \exp\left\{-\frac{1}{2}n\left(\log\left(1-4z^2\right) + 2z\log\left(\frac{1+2z}{1-2z}\right)\right)\right\},$$

where

$$z = \frac{\lambda}{n} < \frac{1}{2}.$$

Now, for $0 \leq x < 1/2$, define

$$f(x) = \log\left(1-4x^2\right) + 2x\log\left(\frac{1+2x}{1-2x}\right) - 4x^2.$$

Then

$$f'(x) = 2\log\left(\frac{1+2x}{1-2x}\right) - 8x.$$

That this is positive can be verified from the power series for the logarithmic term. It follows that, for $x > 0$, $f(x) > f(0) = 0$. Therefore,

$$\log\left(1-4z^2\right) + 2z\log\left(\frac{1+2z}{1-2z}\right) > 4z^2,$$

and

$$2^{-n}\sum < \exp\left(-2z^2\right) = \exp\left(\frac{-2\lambda^2}{n}\right),$$

from which the result follows. □

Proposition 4.10 [Hoeffding's Inequality] *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with zero means and bounded ranges:*

$$a_i \leq Y_i \leq b_i.$$

Then, for any $\eta > 0$, the probability that

$$Y_1 + Y_2 + \ldots + Y_n \geq \eta$$

is at most

$$\exp\left(-\frac{2\eta^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right).$$

□

For a proof of Proposition 4.10, see [27].

Symmetrization proofs

We now prove the symmetrization results, the proofs being similar to those in [35].

Proof of 4.4 Suppose that $\mathbf{x} \in Q$. Then there is $C \in \mathcal{C}$ such that

$$|\mathbf{P}_{\mathbf{x}}(C) - \nu(C)| > \eta.$$

In this case, for $\mathbf{y} \in S^m$,

$$|\mathbf{P}_{\mathbf{y}}(C) - \nu(C)| \leq \frac{\eta}{2} \Longrightarrow |\mathbf{P}_{\mathbf{x}}(C) - \mathbf{P}_{\mathbf{y}}(C)| > \frac{\eta}{2}.$$

Now, the probability (with respect to ν^m) that $|\mathbf{P}_{\mathbf{y}}(C) - \nu(C)| > \eta/2$ is, by Chebyshev's inequality, at most

$$\frac{\nu(C)(1-\nu(C))m}{\left(\frac{\eta m}{2}\right)^2} < \frac{1}{\eta^2 m}.$$

It follows that for $m \geq 2/\eta^2$, $|\mathbf{P}_{\mathbf{y}}(C) - \nu(C)| \not\leq \eta/2$ with probability at least $1/2$. Therefore, applying Fubini's theorem as in the proof of Proposition 3.6,

$$\nu^m(R) \geq \frac{1}{2}\nu^m(Q),$$

and the symmetrization result for Bound One follows. □

Proof of 4.5 Suppose $\mathbf{x} \in V$, so that there is $C \in \mathcal{C}$ with

$$\nu(C) - \mathbf{P}_{\mathbf{x}}(C) > \eta\sqrt{\nu(C)}.$$

Since $\mathbf{P}_{\mathbf{x}}(C) \geq 0$, this implies

$$\nu(C) > \eta^2.$$

Now suppose $m \geq 2/\eta^2$ and $\mathbf{y} \in S^m$ is such that

$$\mathbf{P}_{\mathbf{y}}(C) > \nu(C).$$

Let

$$F = \frac{\mathbf{P}_{\mathbf{y}}(C) - \mathbf{P}_{\mathbf{x}}(C)}{\sqrt{\mathbf{P}_{\mathbf{xy}}(C)}},$$

noting that the denominator is positive since $\mathbf{P}_{\mathrm{y}}(C) > 0$. A simple piece of calculus shows that $F > \eta$.

As Vapnik notes, since $m \geq 2/\eta^2 > 2/\nu(C)$, it follows that with probability at least 1/4 (with respect to ν^m), $\mathbf{P}_{\mathrm{y}}(C) > \nu(C)$. Thus, with probability at least 1/4,

$$\mathbf{P}_{\mathrm{y}}(C) - \mathbf{P}_{\mathrm{x}}(C) > \eta\sqrt{\mathbf{P}_{\mathrm{xy}}(C)}.$$

We therefore have, applying Fubini's theorem,

$$\nu^{2m}(W) \geq \frac{1}{4}\nu^m(V),$$

and the symmetrization result for Bound Two follows. □

Combinatorial bounding proofs

We give two proofs of each of these results.

First proof of 4.7 Let T^i be the one-sided version of R^i,

$$T^i = \left\{\mathrm{xy} \in S^{2m} : (\mathbf{P}_{\mathrm{y}}(A_i) - \mathbf{P}_{\mathrm{x}}(A_i)) > \frac{\eta}{2}\right\},$$

where A_i is as defined on page 72.

Suppose that $\Omega_R^i(\mathrm{z}) \neq 0$. Then there is some σ in Λ such that $\sigma\mathrm{z} \in T^i$. Choose σ so that $\sigma\mathrm{z} \in T^i$ and, writing

$$\sigma\mathrm{z} = \mathrm{xy} = (x_1, \ldots, x_m, y_1, \ldots, y_m),$$

$\mathbf{P}_{\mathrm{y}}(A_i)$ is maximal (and, consequently, $\mathbf{P}_{\mathrm{x}}(A_i)$ is minimal) among all such σ. Let

$$\mathbf{P}_{\mathrm{y}}(A_i) = \frac{r}{m} \text{ and } \mathbf{P}_{\mathrm{x}}(A_i) = \frac{s}{m}.$$

Without loss of generality, we may assume that the r entries of y which belong to A_i are the first r entries of y. Then the s entries of x which belong to A_i must be among the first r entries of x (for, if not, at least one of these entries could be "swapped", contradicting the maximality of r).

Let

$$\omega(\mathrm{z}) = \{\tau \in \Lambda : \tau(\sigma\mathrm{z}) \in T^i\}.$$

Now, because Λ is a group,

$$\begin{aligned}\Omega_R^i(z) &= |\{\tau \in \Lambda : \tau z \in R^i\}| \\ &= 2\,|\{\tau \in \Lambda : \tau z \in T^i\}| \\ &= 2\,|\{\tau \in \Lambda : \tau(\sigma z) \in T^i\}| \\ &= 2\,|\omega(z)|.\end{aligned}$$

There is an obvious one-to-one correspondence between permutations in Λ and subsets of $\{1,...,m\}$; σ maps to the subset

$$S(\sigma) = \{i : 1 \le i \le m,\ \sigma(i) = m + i\}$$

of $\{1,2,\ldots,m\}$ consisting of the positions "swapped" by σ. Suppose that, under this correspondence, $\tau \in \omega(z)$ maps to the subset $\mathcal{T} = S(\tau)$. Then $\mathcal{T}$ can contain any position k such x_k and y_k either both belong to A_i or both do not belong to A_i. Suppose that, in addition, $\mathcal{T}$ contains j of the $r - s$ positions k such that y_k belongs to A_i and x_k does not belong to A_i. Then, since $\tau(\sigma z) \in T^i$, we must have

$$\frac{(r-j)}{m} - \frac{(s+j)}{m} > \frac{\eta}{2}.$$

That is, $j < \delta$, where

$$\delta = \frac{1}{2}(r-s) - \frac{\eta m}{4}.$$

It follows that

$$\Omega_R^i(z) = 2\,|\omega(z)| = 2\,2^s 2^{m-r} \sum_{j<\delta} \binom{r-s}{j}.$$

By Proposition 4.9,

$$\begin{aligned}\frac{\Omega_R^i(z)}{|\Lambda|} &= \frac{\Omega_R^i(z)}{2^m} \\ &< 2\,2^{s-r} 2^{r-s} \exp\left(-\frac{2}{16}\frac{\eta^2 m^2}{(r-s)}\right) \\ &= 2\exp\left(-\frac{\eta^2 m^2}{8(r-s)}\right) \\ &\le 2\exp\left(-\frac{1}{8}\eta^2 m\right),\end{aligned}$$

using

$$(r-s) \leq r \leq m.$$

Therefore

$$\frac{\Omega_R^i(\Lambda, z)}{|\Lambda|} \leq 2\exp\left(-\frac{1}{8}\eta^2 m\right),$$

and the combinatorial bounding result for Bound One follows. □

Second proof of 4.7 We can give a second, more sophisticated, proof of this result using Hoeffding's inequality. Write

$$z = (z_1, z_2, \ldots, z_{2m}).$$

Following Haussler [18], for each $1 \leq j \leq 2m$, we let

$$X_j = \begin{cases} 1 & \text{if } z_j \in A_i; \\ 0 & \text{otherwise.} \end{cases}$$

For $1 \leq j \leq m$, we let Y_j be the random variable

$$Y_j = \begin{cases} X_j - X_{m+j} & \text{with probability } 1/2\ ; \\ X_{m+j} - X_j & \text{with probability } 1/2. \end{cases}$$

Let P be the uniform distribution on Λ. Then,

$$\begin{aligned} \frac{\Omega_R^i(z)}{|\Lambda|} &= 2\frac{1}{|\Lambda|}\left|\left\{\tau \in \Lambda : \frac{1}{m}\sum_{j=1}^{m} X_{\tau^{-1}(m+j)} - \frac{1}{m}\sum_{j=1}^{m} X_{\tau^{-1}(j)} > \frac{\eta}{2}\right\}\right| \\ &= 2P\left\{\sigma \in \Lambda : \sum_{j=1}^{m}\left(X_{\sigma(m+j)} - X_{\sigma(j)}\right) > \frac{\eta m}{2}\right\}. \end{aligned}$$

Now, for a random permutation σ from Λ (chosen according to the uniform distribution on Λ), the elements of $S(\sigma)$ (that is, the swaps present in σ) can be regarded as chosen independently each with a probability $1/2$ of being chosen. Therefore this quantity is twice the probability that

$$\sum_{j=1}^{m} Y_j > \frac{\eta m}{2},$$

and, by Hoeffding's inequality, this is at most

$$2\exp\left(-\frac{\eta^2 m^2}{2\sum_{j=1}^{m}\left(2(X_j - X_{m+j})\right)^2}\right) \leq 2\exp\left(-\frac{1}{8}\eta^2 m\right),$$

as required. □

First proof of 4.8 The arguments are very similar to those given in the preceeding proof. As there, choose $\sigma \in \Lambda$ so that

$$\sigma z = xy = (x_1, \ldots, x_m, y_1, \ldots, y_m) \in W^i,$$

and $\mathbf{P}_y(A_i)$ is maximal (and, consequently, $\mathbf{P}_x(A_i)$ is minimal) among all such σ. Let

$$\mathbf{P}_y(A_i) = \frac{r}{m} \text{ and } \mathbf{P}_x(A_i) = \frac{s}{m}.$$

Again, without loss of generality, assume that the r entries of y which belong to A_i are the first r entries of y and that, consequently, the s entries of x which belong to A_i are among the first r entries of x.

Let

$$\omega(z) = \{\tau \in \Lambda : \tau(\sigma z) \in W^i\}.$$

As above, the fact that Λ is a group implies

$$\begin{aligned} \Omega^i_W(z) &= |\{\tau \in \Lambda : \tau z \in W^i\}| \\ &= |\{\tau \in \Lambda : \tau(\sigma z) \in W^i\}| \\ &= |\omega(z)|. \end{aligned}$$

Under the correspondence S of the last proof, suppose that $\tau \in \omega(z)$ maps to the subset $\mathcal{T} = S(\tau)$. Then $\mathcal{T}$ can contain any position k such that x_k and y_k either both belong to A_i or both do not belong to A_i. Suppose now that, additionally, $\mathcal{T}$ contains j of the $r - s$ positions k such that y_k belongs to A_i and x_k does not belong to A_i. Then, since $\tau(\sigma z) \in W^i$, we must have

$$\frac{(r-j)}{m} - \frac{(s+j)}{m} > \eta\sqrt{\frac{r+s}{2m}}.$$

That is, $j < \delta$, where

$$\delta = \frac{1}{2}(r-s) - \frac{1}{2}\eta\sqrt{\frac{(r+s)m}{2}}.$$

It follows that

$$\Omega^i_W(z) = 2^s 2^{m-r} \sum_{j<\delta} \binom{r-s}{j}.$$

By Proposition 4.9,

$$\begin{aligned}\frac{\Omega^i_W(z)}{|\Lambda|} &= \frac{\Omega^i_W(z)}{2^m} \\ &< 2^{s-r} 2^{r-s} \exp\left(-\frac{1}{4}\frac{\eta^2(r+s)m}{(r-s)}\right) \\ &\leq \exp\left(-\frac{1}{4}\eta^2 m\right).\end{aligned}$$

The symmetrization result follows. □

Second proof of 4.8 Again, we can give a second proof using Hoeffding's inequality. In this case, with the same definitions of the variables X_i and Y_i, for $1 \leq i \leq m$, we have

$$\begin{aligned}&\frac{\Omega^i_W(z)}{|\Lambda|} \\ &= \frac{1}{|\Lambda|}\left|\left\{\tau \in \Lambda : \frac{1}{m}\sum_{j=1}^{m} X_{\tau^{-1}(m+j)} - \frac{1}{m}\sum_{j=1}^{m} X_{\tau^{-1}(j)} > \eta\left(\frac{1}{2m}\sum_{j=1}^{2m} X_j\right)^{1/2}\right\}\right| \\ &= P\left\{\sigma \in \Lambda : \sum_{j=1}^{m}\left(X_{\sigma(m+j)} - X_{\sigma(j)}\right) > \eta\left(\frac{m}{2}\sum_{j=1}^{2m} X_j\right)^{1/2}\right\} \\ &\leq \exp\left(-\frac{2\eta^2 m \sum_{j=1}^{2m} X_j}{2\sum_{j=1}^{m}(2(X_j - X_{m+j}))^2}\right).\end{aligned}$$

Observing that

$$\sum_{j=1}^{m}(X_j - X_{m+j})^2 \leq \sum_{i=1}^{2m} X_j,$$

we have

$$\frac{\Omega^i_W(z)}{|\Lambda|} \leq \exp\left(-\frac{1}{4}\eta^2 m\right),$$

as required. □

4.4 A Result of Chapter 3

In the previous chapter, Corollary 3.9, we gave a proof of essentially the following result.

Theorem 4.11 *With the definitions of this chapter, for* $\mathrm{x} \in S^m$, *let* $\mathcal{C}[\mathrm{x}]$ *denote the subclass of* $\mathcal{C}$ *consisting of those events* A *for which* $\mathbf{P}_{\mathrm{x}}(A) = 0$. *Then, for any* $\eta > 0$, *and for any positive integer* $m \geq 8/\eta$,

$$\nu^m \left\{ \mathrm{x} \in S^m : \sup_{\mathcal{C}[\mathrm{x}]} \nu(A) > \eta \right\} \leq 2\,\Pi_{\mathcal{C}}(2m)\, 2^{-\eta m/2}.$$

□

We have seen that this result is very useful in learnability theory. The same from of result, with slightly weaker constants, can be obtained directly from Bound Two.

If $\mathrm{x} \in S^m$ and $A \in \mathcal{C}[\mathrm{x}]$ is such that $\nu(A) > \eta$ then, in particular, since $\mathbf{P}_{\mathrm{x}}(A) = 0$,

$$\frac{\nu(A) - \mathbf{P}_{\mathrm{x}}(A)}{\sqrt{\nu(A)}} = \sqrt{\nu(A)} > \sqrt{\eta}.$$

By Bound Two, this can occur with probability at most

$$4\,\Pi_{\mathcal{C}}(2m) \exp\left(-\frac{1}{4}\eta m\right).$$

If we apply Bound One, we obtain a weaker result with η^2 in the exponent.

The techniques of this chapter can easily be applied to obtain a simple direct proof of Theorem 4.11. The symmetrization result is the same as that for Bound One, and the combinatorial bounding argument is particularly simple in this case; no binomial sums are required. We sketch the proof below.

Proof of 4.11 Let

$$Q = \left\{ \mathrm{x} \in S^m : \sup_{\mathcal{C}[\mathrm{x}]} \nu(A) > \eta \right\}.$$

Let

$$T = \left\{ \mathrm{xy} \in S^{2m} : \sup_{C[\mathrm{x}]} \mathbf{P}_{\mathrm{y}}(A) > \frac{\eta}{2} \right\}.$$

Fix $\mathrm{z} \in S^{2m}$, and suppose that $\{A_1, \ldots, A_t\}$ is a complete set of distinct representatives of $\mathcal{C}$ for z. For each i between 1 and t, let

$$T^i = \left\{ \mathrm{xy} \in S^{2m} : \mathbf{P}_{\mathrm{x}}(A_i) = 0,\ \mathbf{P}_{\mathrm{y}}(A_i) > \frac{\eta}{2} \right\}.$$

Then, using Chebyshev's inequality and Fubini's theorem, as in chapter 3, and using an argument similar to the symmetrization proof for Bound One,

$$\nu^m(Q) \leq 2\,\nu^{2m}(T).$$

Suppose that

$$\sigma \mathrm{z} = \mathrm{xy} = (x_1, \ldots, x_m, y_1, \ldots, y_m) \in T^i$$

and that

$$\mathbf{P}_{\mathrm{y}}(A_i) = \frac{r}{m} > \frac{\eta}{2}.$$

Let $\tau \in \Lambda$ be such that $\tau(\mathrm{xy}) \in T^i$. It is easy to see that only for k such that $y_k \notin A_i$ can τ be such that $\tau(m+k) = k$. Therefore, as before,

$$\begin{aligned} \frac{\Omega_T^i(\mathrm{z})}{|\Lambda|} &= \frac{1}{2^m} \left| \{ \tau \in \Lambda : \tau(\sigma \mathrm{z}) \in T^i \} \right| \\ &= \frac{1}{2^m} 2^{m-r} \\ &= 2^{-r}, \end{aligned}$$

and this is at most $2^{-\eta m/2}$ since $r > \eta m/2$. The result follows. □

Chapter 5
Stochastic Concepts

5.1 Introduction

In this chapter, we discuss stochastic concepts [11]. Until now, we have considered the learnability of $\{0,1\}$-valued functions (or, equivalently, sets). One reason for this is that we are attempting to approximate to an underlying target concept by a $\{0,1\}$-function, and if this approximation is to be guaranteed to any arbitrary degree of closeness, the concept itself must be such a function. Thus the target concept is sharply defined; a given input is either a positive example or a negative example of the concept and there is no ambiguity. There are many reasons why it is unrealistic to assume that the object being learned is a function. In many real learning situations there may be some inputs which should not be classified as definitely positive examples or definitely negative examples, but rather as somewhat positive and somewhat negative (in some sense). For example, it may not be clear what classification should be given to points very close to the edge of an object in a pattern recognition problem.

Even when the concept to be learned is indeed well-defined (that is, a function), the training examples may be randomly misclassified to some degree during the training procedure. Theoretically, one may like to consider this as learning from a faulty teacher. In practice, it may be due to some electrical "noise" in a computer implementation of a learning algorithm. We briefly discuss how consideration of stochastic concepts has proved useful in studying learning in the presence of such classification errors [20].

Another situation in which the idea of stochastic concepts is useful is that in which the hypothesis output by the learning algorithm is not consistent with all of the training examples that have been presented during the training procedure, but rather only with at least a certain fraction of them. This is a realistic situation; even when there is no noise and the target concept is well-defined, it is a great restriction on the learning algorithm to stipulate that after training, it output a hypothesis consistent with all of the training sample.

5.2 Stochastic Concepts

Stochastic concepts

Stochastic concepts are defined in such a way that, with respect to a stochastic concept, an input need not be either a positive example or a negative example but, rather, has a certain probability as a positive example and a certain probability as a negative example. Thus it is possible that a particular input may be presented sometimes as a positive example and sometimes as a negative example during training.

As before, X denotes the input space, which is finite, countable or Euclidean. Σ is a σ-algebra of subsets of X which in the case of countable X is the power set of X and in the case of Euclidean X is the induced Borel σ-algebra. Throughout, S will denote the cartesian product $S = X \times \{0,1\}$ and Φ the product σ-algebra $\Sigma \times 2^{\{0,1\}}$ of subsets of S. We make the following definition, following [11].

Definition 5.1 *A stochastic concept on X is a probability measure ν defined on the σ-algebra $\Phi = \Sigma \times 2^{\{0,1\}}$ of subsets of $X \times \{0,1\}$.* □

Deterministic concepts

If the definition of stochastic concept is to be a sensible one, it must generalize the previous framework in which a concept is regarded as a measurable function from X to $\{0,1\}$ and there is an underlying probability measure μ defined on the σ-algebra Σ. This is indeed the case. We require the following lemma.

Lemma 5.2 *The σ-algebra $\Phi = \Sigma \times 2^{\{0,1\}}$ consists precisely of all sets of the form*

$$(A_0 \times \{0\}) \cup (A_1 \times \{1\}),$$

with $A_0, A_1 \in \Sigma$.

Proof Let $\mathcal{A}$ denote the collection of sets

$$\mathcal{A} = \{(A_0 \times \{0\}) \cup (A_1 \times \{1\}) : A_0, A_1 \in \Sigma\}.$$

Then it is easily verified that $\mathcal{A}$ is a σ-algebra containing all product rectangles. Therefore $\mathcal{A} \supseteq \Phi$. But the reverse inclusion is clear, and therefore $\mathcal{A} = \Phi$. □

Consider any measurable function c from X to $\{0,1\}$ and any probability measure μ on (X, Σ). We can define a measure $\nu = \nu(c, \mu)$ on Φ by defining ν on an arbitrary member of Φ as follows:

$$\nu((A_0 \times \{0\}) \cup (A_1 \times \{1\})) = \mu\left(c^{-1}(0) \cap A_0\right) + \mu\left(c^{-1}(1) \cap A_1\right).$$

Then ν is easily seen to be a probability measure on Φ which represents the pair (c, μ) in the following way.

Proposition 5.3 *For any $A \in \Sigma$, let*

$$A_c = \{(x, c(x)) : x \in A\}$$

and let

$$\hat{A}_c = (A \times \{0,1\}) \setminus A_c = \{(x,y) : x \in A,\ y \neq c(x)\}.$$

Then both these sets belong to the σ-algebra Φ and if $\nu = \nu(c, \mu)$ is the measure defined above,

$$\nu(A_c) = \mu(A), \quad \nu(\hat{A}_c) = 0.$$

Proof Consider the set A_c. We have

$$\begin{aligned} A_c &= \{(x, c(x)) : x \in A\} \\ &= \{(x,0) : x \in A,\ c(x) = 0\} \cup \{(x,1) : x \in A,\ c(x) = 1\} \\ &= \left(c^{-1}(0) \cap A\right) \times \{0\} \cup \left(c^{-1}(1) \cap A\right) \times \{1\}, \end{aligned}$$

which is measurable. Further,

$$\begin{aligned} \nu(A_c) &= \nu\left(\left(c^{-1}(0) \cap A\right) \times \{0\} \cup \left(c^{-1}(1) \cap A\right) \times \{1\}\right) \\ &= \mu\left(c^{-1}(0) \cap A\right) + \mu\left(c^{-1}(1) \cap A\right) \\ &= \mu\left(\left(c^{-1}(0) \cap A\right) \cup \left(c^{-1}(1) \cap A\right)\right) \\ &= \mu(A). \end{aligned}$$

Now, $\hat{A}_c = (A \times \{0,1\}) \setminus A_c$ is measurable since $A \times \{0,1\}$ and A_c are measurable. Further,

$$\begin{aligned} \nu(A \times \{0,1\}) &= \nu(A \times \{0\} \cup A \times \{1\}) \\ &= \mu\left(c^{-1}(0) \cap A\right) + \mu\left(c^{-1}(1) \cap A\right) \\ &= \mu(A). \end{aligned}$$

It follows that

$$\nu(\hat{A}_c) = \nu\left((A \times \{0,1\}) \setminus A_c\right) = \nu(A \times \{0,1\}) - \nu(A_c) = 0.$$

□

Thus $\nu(c, \mu)$ represents the target concept c together with the underlying distribution μ. We call the stochastic concept $\nu(c, \mu)$ the *deterministic concept representing c and μ.*

5.3 Approximating Stochastic Concepts by Functions

In this section, we shall consider the approximation of stochastic concepts on X by a space of measurable functions from X to $\{0,1\}$. The framework is as follows: We have a set H, the *hypothesis space*, of measurable functions from X to $\{0,1\}$ and we are to approximate the target stochastic concept ν by a hypothesis from H.

The error of a hypothesis

Suppose that ν is some target stochastic concept on X. For any $h \in H$, the set

$$\{(x,a) : a \neq h(x)\}$$

is a Φ-measurable set. To see this, observe that

$$\begin{aligned}\{(x,a) : a \neq h(x)\} &= \{(x,1) : h(x) = 0\} \cup \{(x,0) : h(x) = 1\} \\ &= \left(h^{-1}(0) \times \{1\}\right) \cup \left(h^{-1}(1) \times \{0\}\right),\end{aligned}$$

which, as the union of two measurable sets, is measurable. We can therefore define the *actual error* (with respect to ν) of $h \in H$ to be

$$\mathrm{er}_\nu(h) = \nu\{(x,a) : a \neq h(x)\}.$$

Notice that if $\nu = \nu(c,\mu)$ is the deterministic concept representing (c,μ), then

$$\begin{aligned}\mathrm{er}_\nu(h) &= \nu\{(x,a) : a \neq h(x)\} \\ &= \nu\left(\left(h^{-1}(1) \times \{0\}\right) \cup \left(h^{-1}(1) \times \{1\}\right)\right).\end{aligned}$$

Now, h is measurable and so $h^{-1}(0), h^{-1}(1) \in \Sigma$. Therefore

$$\begin{aligned}\mathrm{er}_\nu(h) &= \mu\left(h^{-1}(1) \cap c^{-1}(0)\right) + \mu\left(h^{-1}(0) \cap c^{-1}(1)\right) \\ &= \mu\left(\{x : c(x) \neq h(x)\}\right).\end{aligned}$$

This coincides with the previous definition of the actual error of h with respect to c when the underlying probability measure is μ.

For a subset F of H, we define the *haziness of* F with respect to ν to be

$$\mathrm{haz}_\nu(F) = \sup\{\mathrm{er}_\nu(f) : f \in F\}.$$

A *sample* of length m of ν is a sequence $\mathbf{y}$ of m points of S, randomly drawn according to the distribution ν. We regard $\mathbf{y}$ as an element of the product space $(X \times \{0,1\})^m$. For $h \in H$, the *observed error* of h on sample

$$\mathbf{y} = ((x_1, a_1), \ldots, (x_m, a_m))$$

is defined to be

$$\mathrm{er}_{\mathbf{y}}(h) = \frac{1}{m} |\{i : h(x_i) \neq a_i\}| .$$

Clearly, it is not possible to approximate a stochastic concept ν arbitrarily closely with a function from H unless ν is a deterministic concept representing a hypothesis from H. However, we should like to be able to give a guarantee that, with high probability, any hypothesis h from H which has small observed error on a random sample of ν of sufficient length is indeed a good approximation to ν. By this, we mean that, with probability at least $1 - \epsilon$, for a randomly chosen point $(x, a) \in S$, $h(x) = a$; that is, h has actual error less than ϵ with respect to ν.

The proofs that such a guarantee can be given for the case of finite hypothesis spaces and for hypothesis spaces of finite VC dimension form the remainder of this section. For finite hypothesis spaces, we use an estimation for the tail of the binomial distribution, while for hypothesis spaces of finite VC dimension, as in [11] we use the powerful theory of Chapter 4, obtaining sample-size bounds which improve the best previously known.

Finite hypothesis spaces

We consider here the approximation of a stochastic concept on X by a finite set of $\{0,1\}$-valued functions defined on X. We first state a result of Angluin and Valiant [3] which provides a useful bound on the tail of a binomial series. For a proof of this result, see [21].

Lemma 5.4 For any $0 < \beta < 1$, for any $0 < p < 1$ and for any positive integer n,

$$\sum_{i=0}^{\lfloor (1-\beta)np \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \leq \exp\left(-\frac{\beta^2 np}{2}\right).$$

□

Using this, we can prove the following [11].

Theorem 5.5 *Let H be a finite hypothesis space of $\{0,1\}$-valued functions defined on an input space X. Let ν be any probability measure on $S = X \times \{0,1\}$ (that is, a stochastic concept on X), let $0 < \epsilon < 1$ and $0 < \gamma \leq 1$. Then the probability (with respect to ν^m) that, for $\mathbf{y} \in S^m$, there is some hypothesis from H such that*

$$\mathrm{er}_\nu(h) > \epsilon \quad \text{and} \quad \mathrm{er}_\mathbf{y}(h) \leq (1-\gamma)\mathrm{er}_\nu(h)$$

is less than

$$|H| \exp\left(-\frac{1}{2}\gamma^2 \epsilon m\right).$$

Proof For any $h \in H$ with $\mathrm{er}_\nu(h) = \epsilon_h > \epsilon$ and for any $0 < \gamma \leq 1$, we have

$$\begin{aligned}
\nu^m \{\mathbf{y} \in S^m : \mathrm{er}_\mathbf{y}(h) \leq (1-\gamma)\epsilon_h\} &= \sum_{i=0}^{\lfloor (1-\gamma)\epsilon_h m \rfloor} \binom{m}{i} \epsilon_h^i (1-\epsilon_h)^{m-i} \\
&\leq \exp\left(-\frac{1}{2}\gamma^2 \epsilon_h m\right) \\
&< \exp\left(-\frac{1}{2}\gamma^2 \epsilon m\right),
\end{aligned}$$

by Lemma 5.4. The probability that there is some h with $\mathrm{er}_\nu(h) > \epsilon$ and $\mathrm{er}_\mathbf{y}(h) \leq (1-\gamma)\mathrm{er}_\nu(h)$ is bounded by $|H|$ times this quantity, and the result follows. □

Corollary 5.6 *Let $0 < \epsilon, \delta < 1$ and $0 < \gamma \leq 1$ and let ν be any probability measure on $S = X \times \{0,1\}$. If H is a finite hypothesis space then there is an integer $m_0 = m_0(\epsilon, \delta, \gamma)$ such that if $m \geq m_0$ then, for $\mathbf{y} \in S^m$, with probability at least $1 - \delta$ (with respect to ν^m),*

$$\mathrm{er}_{\mathbf{y}}(h) \leq (1-\gamma)\epsilon \Longrightarrow \mathrm{er}_{\nu}(h) \leq \epsilon.$$

A suitable value of m_0 is

$$m_0(\epsilon, \delta, \gamma) = \left\lceil \frac{2}{\gamma^2 \epsilon} \log\left(\frac{|H|}{\delta}\right) \right\rceil,$$

where log *denotes natural logarithm.*

Proof We have

$$\begin{aligned} & |H| \exp\left(-\frac{1}{2}\gamma^2 \epsilon m\right) \leq \delta \\ \Longleftrightarrow\ & \log|H| - \frac{1}{2}\gamma^2 \epsilon m \leq \log \delta \\ \Longleftrightarrow\ & m \geq \frac{2}{\gamma^2 \epsilon} \log\left(\frac{|H|}{\delta}\right). \end{aligned}$$

□

This result shows that, in a finite hypothesis space, if a hypothesis is a good approximation to the target stochastic concept on a large enough sample of the concept (the sufficient size being independent of the target stochastic concept) then it is probably a good approximation to the target on the whole input space (where "probably" and "good approximation" have the usual technical meanings).

Hypothesis spaces of finite VC dimension

We can apply Bound Two of Chapter 4 to the approximation of stochastic concepts by a hypothesis space H of finite VC dimension.

Let S be $X \times \{0,1\}$ and suppose that H is a hypothesis space of measurable functions from X to $\{0,1\}$. For any hypothesis h from H, we define the *error set* E_h of h by

$$E_h = \{(x,a) \in S : a \neq h(x)\}.$$

Let $\mathcal{C}$ be the collection of all error sets of hypotheses in H. That is,

$$\mathcal{C} = \{E_h : h \in H\}.$$

The following result (an extension of a result from [11]) shows that $\mathcal{C}$ (as a class of subsets of S) has the same growth function as H (as a class of functions from X to $\{0,1\}$):

Lemma 5.7 *For any positive integer m, $\Pi_{\mathcal{C}}(m) = \Pi_H(m)$. In particular, $\mathcal{C}$ and H have the same VC dimension.*

Proof Let $\mathbf{y} = ((x_1,a_1),\ldots,(x_m,a_m)) \in S^m$, and let

$$I(\mathbf{y}) = \{(x_i,a_i) : 1 \leq i \leq m\}.$$

Suppose that $h, g \in H$. Then

$$\begin{aligned} & E_h \cap I(\mathbf{y}) = E_g \cap I(\mathbf{y}) = \{(x_i,a_i) : i \in J\} \\ & \Longrightarrow h(x_j) = g(x_j) \neq a_j \ (j \in J) \text{ and } h(x_i) = g(x_i) = a_i \ (i \notin J) \\ & \Longrightarrow h(x_i) = g(x_i) \ (1 \leq i \leq m). \end{aligned}$$

Hence

$$E_h \cap I(\mathbf{y}) = E_g \cap I(\mathbf{y}) \Longleftrightarrow h(x_i) = g(x_i) \ (1 \leq i \leq m).$$

Therefore the number of distinct sets of the form $C \cap I(\mathbf{y})$ where $C \in \mathcal{C}$ is equal to the number of distinct vectors of the form $(h(x_1),\ldots,h(x_m))$ where $h \in H$. Thus, for any $\mathbf{y} \in S^m$ there is $\mathbf{x} \in X^m$ such that

$$\Pi_{\mathcal{C}}(\mathbf{y}) = \Pi_H(\mathbf{x})$$

and, consequently, $\Pi_{\mathcal{C}}(m) \leq \Pi_H(m)$. Conversely, for any $\mathbf{x} \in X^m$, let $\mathbf{y}$ be as above, with any choice of the $a_i (1 \leq i \leq m)$. The above analysis shows that $\Pi_{\mathcal{C}}(\mathbf{y}) = \Pi_H(\mathbf{x})$ and therefore we obtain the reverse inequality $\Pi_{\mathcal{C}}(m) \geq \Pi_H(m)$. Thus $\Pi_{\mathcal{C}}(m) = \Pi_H(m)$. It follows immediately that $\mathcal{C}$ and H have the same VC dimension. □

We therefore have the following result, applying Bound Two to the family $\mathcal{C}$.

Theorem 5.8 *Let H be a hypothesis space of $\{0,1\}$-valued functions defined on an input space X. Let ν be any probability measure on $S = X \times \{0,1\}$ (that is, a stochastic concept on X), let $0 < \epsilon < 1$ and let $0 < \gamma \leq 1$. Then the probability (with respect to the product measure ν^m) that, for $\mathbf{y} \in S^m$, there is some hypothesis from H such that*

$$\mathrm{er}_\nu(h) > \epsilon \quad \text{and} \quad \mathrm{er}_{\mathbf{y}}(h) \leq (1-\gamma)\mathrm{er}_\nu(h)$$

is at most

$$4\,\Pi_H(2m)\exp\left(-\frac{1}{4}\gamma^2\epsilon m\right).$$

Proof We apply Bound Two of Chapter 4. As suggested above, we take $S = X \times \{0,1\}$ and $\mathcal{C}$ to be the collection of error sets of the hypotheses from H. For any $E_h \in \mathcal{C}$, the relative frequency of occurence of event E_h on sample

$$\mathbf{y} = (y_1, \ldots, y_m) = ((x_1, a_1), \ldots, (x_m, a_m)) \in S^m$$

is

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(E_h) &= \frac{1}{m}\left|\{i : (x_i, a_i) \in E_h\}\right| \\ &= \frac{1}{m}\left|\{i : h(x_i) \neq a_i\}\right| \\ &= \mathrm{er}_{\mathbf{y}}(h),\end{aligned}$$

the observed error of h on the sample $\mathbf{y}$. Further,

$$\nu(E_h) = \nu\{(x,a) : h(x) \neq a\} = \mathrm{er}_\nu(h),$$

the actual error of h with respect to ν. Now, if h is such that $\mathrm{er}_\nu(h) = \epsilon_h > \epsilon$ and $\mathrm{er}_{\mathbf{y}}(h) \leq (1-\gamma)\epsilon_h$ then

$$\mathrm{er}_\nu(h) - \mathrm{er}_{\mathbf{y}}(h) \geq \epsilon_h - (1-\gamma)\epsilon_h = \gamma\epsilon_h,$$

and hence

$$\frac{\mathrm{er}_\nu(h) - \mathrm{er}_{\mathbf{y}}(h)}{\sqrt{\mathrm{er}_\nu(h)}} \geq \frac{\gamma\epsilon_h}{\sqrt{\epsilon_h}} = \gamma\sqrt{\epsilon_h} > \gamma\sqrt{\epsilon}.$$

That is,

$$\frac{\nu(E_h) - \mathbf{P}_{\mathbf{y}}(E_h)}{\sqrt{\nu(E_h)}} > \gamma\sqrt{\epsilon}.$$

By Bound Two, the ν^m-measure of the set of $\mathbf{y} \in S^m$ for which such an h exists is at most

$$4\,\Pi_C(2m)\exp\left(-\frac{1}{4}(\gamma\sqrt{\epsilon})^2 m\right) = 4\,\Pi_H(2m)\exp\left(-\frac{1}{4}\gamma^2\epsilon m\right),$$

where we have used Lemma 5.7. □

Corollary 5.9 *Let $0 < \epsilon, \delta < 1$ and $0 < \gamma \leq 1$ and let ν be any distribution on $S = X \times \{0,1\}$. If H has finite VC dimension d, then there is an integer $m_0 = m_0(\epsilon, \delta, \gamma)$ such that if $m \geq m_0$ then, for $\mathbf{y} \in S^m$, with probability at least $1 - \delta$ (with respect to the product measure ν^m),*

$$\mathrm{er}_{\mathbf{y}}(h) \leq (1-\gamma)\epsilon \Longrightarrow \mathrm{er}_\nu(h) \leq \epsilon.$$

A suitable value of m_0 is

$$m_0(\epsilon, \delta, \gamma) = \left\lceil \frac{1}{\gamma^2\epsilon(1-\sqrt{\epsilon})}\left(4\log\left(\frac{4}{\delta}\right) + 6d\log\left(\frac{4}{\gamma^{2/3}\epsilon}\right)\right)\right\rceil,$$

where log *denotes natural logarithm.*

Proof The proof uses Sauer's inequality and Lemma 3.14; for any $\alpha, x > 0$, $\log x \leq (-\log\alpha - 1) + \alpha x$. H has finite VC dimension d and therefore, for $2m \geq d$, by Sauer's Lemma,

$$\Pi_H(2m) < \left(\frac{2em}{d}\right)^d.$$

We show that if $m \geq m_0$, then

$$4\left(\frac{2em}{d}\right)^d \exp\left(-\frac{1}{4}\gamma^2\epsilon m\right) \leq \delta,$$

from which the result follows.

Now,

$$4\left(\frac{2em}{d}\right)^d \exp\left(-\frac{1}{4}\gamma^2\epsilon m\right) \leq \delta$$

$$\iff \log 4 + d\log 2 + d + d\log m - d\log d - \frac{1}{4}\gamma^2\epsilon m \leq \log\delta$$

$$\iff \frac{1}{4}\gamma^2\epsilon m \geq \log\left(\frac{4}{\delta}\right) + d\log 2 - d\log d + d + d\log m.$$

Let $\alpha = q\gamma^2\epsilon/4d$ in Lemma 3.14, where $0 < q < 1$ is to be chosen. We have

$$\begin{aligned} d\log m &\leq d\left(\log\left(\frac{4d}{\gamma^2\epsilon q}\right) - 1\right) + \frac{\gamma^2\epsilon q}{4}m \\ &= d\log 4d - d\log\gamma^2 - d\log\epsilon - d\log q - d + \frac{\gamma^2\epsilon q}{4}m. \end{aligned}$$

Therefore, it suffices to have

$$\begin{aligned} \frac{1}{4}\gamma^2\epsilon m(1-q) &\geq \log\left(\frac{4}{\delta}\right) + d\log 2 - d\log d + d \\ &\quad + d\log 4d - d\log\gamma^2 - d\log\epsilon - d\log q - d \\ &= \log\left(\frac{4}{\delta}\right) + d\log\left(\frac{8}{\epsilon}\right) + d\log\left(\frac{1}{\gamma^2}\right) - d\log q. \end{aligned}$$

So a sufficient sample size is

$$\left\lceil \frac{4}{\gamma^2\epsilon(1-q)}\left(\log\left(\frac{4}{\delta}\right) + d\log\left(\frac{8}{\epsilon}\right) + d\log\left(\frac{1}{\gamma^2}\right) + d\log\left(\frac{1}{q}\right)\right)\right\rceil.$$

Choosing $q = \sqrt{\epsilon}$ gives the result. □

This sample-size bound is better than that previously obtained [11] for this problem.

We can prove similar results using Bound One of Chapter 4. However, the resulting sample-size involves the reciprocal of ϵ^2 rather than that of ϵ.

5.4 Classification Noise and Semi-Consistent Learning

In this section, we briefly describe how the above results on approximating stochastic concepts can be applied to two of the problems mentioned earlier: learning in the presence of classification errors and learning when the hypothesis output by the learner need not be consistent with all of the training sample.

Classification error

There are many types of error that can occur during a practical implementation of a learning algorithm [32]. We do not attempt to describe all of these, but we briefly describe a particular type of error or noise, which is to be thought of as a random misclassification of the examples presented during the training part of the learning process. This may be due to some degree of electrical "noise" or due to a fault with the teacher or oracle presenting the examples to the learner [2, 20]. Random classification error has been described by Angluin and Laird [2] as follows: Suppose that the target concept is c and the probability distribution on the input space X is μ. As before, during training, the training inputs are randomly drawn according to μ. Suppose that input x has been chosen and that, without loss, $c(x) = 1$. A biased coin is then thrown, and with probability γ_x, x is presented as a negative example of the target concept (that is, x is labelled with 0) and with probability $1 - \gamma_x$, x is correctly presented as a positive example of the concept. Regarding examples as points from $S = X \times \{0,1\}$, this process can be modelled by considering the examples as being chosen randomly from a distribution ν on $S = X \times \{0,1\}$. In this context, ν is to be thought of as a *corrupted* version of the pair (c, μ) (or rather, as a corrupted version of the deterministic concept $\nu(c, \mu)$). It is generally assumed that the γ_x are uniformly bounded by some constant γ.

Suppose now that we want the learner not to be able to learn the target concept c, but to be able to approximate to the distribution ν with a function from a hypothesis space of finite VC dimension. That is, we wish to be able to find a hypothesis from H which provides a good approximation to the stochastic

concept ν. The above theory is immediately applicable. If one finds *any* hypothesis which has observed error at most $(1-\gamma)\epsilon$ on a random sample of ν of sufficient length, then with high probability $1-\delta$, the hypothesis has actual error at most ϵ with respect to ν. Corollary 5.9 provides an upper bound on how large a sample will suffice.

Semi-consistent learning

As mentioned before, a learning algorithm need not produce a hypothesis consistent with all the training examples presented to it. This is clearly true if the examples are examples of a stochastic concept, where during training a particular input may have been presented as both a positive example and a negative example of the target concept. However, even if the concept to be learned is truly deterministic (when the teacher is faultless, the concept is well-defined, and there is no noise), it is perhaps unreasonable to demand that the learner output a fully consistent hypothesis. If the learner outputs a hypothesis from a hypothesis space H of finite VC dimension which agrees with at least a sufficiently large fraction of the training examples, then it seems that learnability to some accuracy should still be guaranteed (albeit with a larger sample-size than for fully-consistent learnability). This is indeed the case, as the preceeding theory shows. Suppose that on presentation of a sample of length m of a (stochastic) concept ν, the learner has produced a hypothesis which has observed error at most $(1-\gamma)\epsilon$ with respect to ν on the sample. That is, the hypothesis correctly classifies at least a fraction $1-(1-\gamma)\epsilon$ of the sample. Corollary 5.9 shows that there is a sufficient sample-size m_0 for which it can be guaranteed that if $m \geq m_0$ then this hypothesis, with high probability $1-\delta$, has error less than ϵ with respect to ν. If ν is the deterministic concept $\nu = \nu(c, \mu)$, this is equivalent to saying that the hypothesis has error less than ϵ with respect to c and μ. Notice that, here, to guarantee that the hypothesis has error less than ϵ with respect to the target (stochastic) concept, we have to demand that the error of the hypothesis on the random sample be less than some definite fraction of ϵ.

Chapter 6
Non-Uniform Learnability

6.1 The Notion of Non-Uniform Learnability

Introduction

In many realistic learning problems, the distribution on the input space is fixed but unknown. This is the primary reason for proving learnability results and finding sufficient sample-sizes which are independent of the distribution; results that are independent of the distribution certainly hold for any particular distribution. If something *is* known of the distribution, it may be possible to say more, proving learnability-type results and finding sample-size bounds even when the hypothesis space is not of finite VC dimension.

In order to introduce non-uniform learnability, we can start from basics by considering the learnability of a *particular* concept c from a hypothesis space H, with respect to a *particular* probability measure μ on the input space X. We say that c is *μ-learnable* in H if given any $\epsilon, \delta \in (0,1)$, there is an integer $m_0 = m_0(\epsilon, \delta)$ such that for all $m \geq m_0$,

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}, c]) > \epsilon\} < \delta.$$

The definition of learnability of H is obtained by stipulating that every hypothesis from H be μ-learnable with respect to every probability measure μ on X, with a sufficient sample-size m_0 which is independent of both the target concept and the distribution μ. The idea of non-uniform learnability is to allow weaker conditions than this, allowing m_0 to depend in various ways on the target concept and the distribution. We make this precise below.

Uniformity parameters for learnability

Recall the definition of learnability of a hypothesis space H (Definition 1.7). H is learnable if for any accuracy parameter ϵ, any confidence parameter δ, any target concept $c \in H$ and any probability measure μ on X, there is a sample-size m_0, which is a function of ϵ and δ alone, such that the following holds: With probability at least $1 - \delta$, if some hypothesis h is consistent with c on more than m_0 inputs chosen randomly according to the distribution μ, then h has actual error less than ϵ. As emphasised earlier, the value of m_0 must depend on neither the target concept c nor the distribution (probability measure) μ. Similarly, the sample-size bounds in Chapter 5 on approximating stochastic concepts by functions from a hypothesis space depend only on the accuracy and confidence parameters and not on the (stochastic) target concept.

These are very stringent requirements. Indeed, we have seen that these forms of learnability only hold when the hypothesis space has finite VC dimension. Following Ben-David *et al* [7], and weakening these requirements by allowing some degrees of dependence on target concept and distribution, we obtain four definitions that are of the standard learnability format, but which are parameterized by uniformity conditions.

Definitions 6.1 *For a hypothesis space H, we say that*

(a) H is $L(c, \mu)$,
(b) H is $L(c, \overline{\mu})$,
(c) H is $L(\overline{c}, \mu)$,
(d) H is $L(\overline{c}, \overline{\mu})$

if for all $c \in H$ and for all probability measures μ on X, there is an integer m_0 such that for all $m \geq m_0$,

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < \delta,$$

(that is, any $c \in H$ is μ-learnable), and

(a) m_0 *is a function of* ϵ, δ *only,*

(b) m_0 *is a function of* ϵ, δ *and* μ *only,*

(c) m_0 *is a function of* ϵ, δ *and* c *only,*

(d) m_0 *is a function of* ϵ, δ, c *and* μ.

□

We shall, for example, talk of the class $L(\overline{c}, \overline{\mu})$ of hypotheses spaces, meaning the set of all hypothesis spaces which are $L(\overline{c}, \overline{\mu})$. In particular, the class $L(c, \mu)$ is precisely the set of learnable hypothesis spaces.

6.2 Distribution-Independent Learnability

Ben-David *et al* [7] have shown that the class $L(\overline{c}, \mu)$ of hypothesis spaces which are learnable uniformly over all distributions on the input space is precisely the class $L(c, \mu)$ of hypothesis spaces learnable uniformly over all concepts and all distributions (and is, therefore, the class of hypothesis spaces of finite VC dimension). The proof of the result follows immediately from a closer analysis of the proof of Theorem 3.16, but we give a direct proof for completeness.

Theorem 6.2 *If* H *is* $L(\overline{c}, \mu)$ *then* H *has finite VC dimension.*

Proof Suppose that H has infinite VC dimension. Let $c \in H$ be a fixed target concept, and let m be any positive integer. We show that there is some probability measure μ on X such that

$$\mu^m \left\{ \mathrm{x} \in X^m : \mathrm{haz}_\mu \left(H[\mathrm{x}] \right) \geq \frac{1}{2} \right\} = 1,$$

and, consequently, H is not $L(\overline{c}, \mu)$.

Since H has infinite VC dimension, there is some $Y \subseteq X$ with $|Y| = 2m$ such that Y is shattered by H. Note that the $2m$ elements of Y are distinct. Define μ on X by defining, for a measurable subset A of X,

$$\mu(A) = \frac{1}{2m} |A \cap Y| .$$

Thus μ is the probability measure that is uniform on Y and zero elsewhere.

Suppose that $\mathrm{x} = (x_1, \ldots, x_m) \in Y^m$, and let

$$F = \{x_i : 1 \leq i \leq m\} \subseteq Y.$$

Since Y is shattered by H, there is $h \in H$ which agrees with c on every element of F and disagrees with c on every element of $Y \setminus F$. Then $h \in H[\mathrm{x}]$ and

$$\mathrm{er}_\mu(h) = \mu(Y \setminus F) \geq \frac{m}{2m} = \frac{1}{2}.$$

This shows that

$$Y^m \subseteq \left\{ \mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) \not\geq \frac{1}{2} \right\},$$

and therefore

$$\mu^m \left\{ \mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) \geq \frac{1}{2} \right\} \geq \mu^m(Y^m) = 1.$$

□

Corollary 6.3 *H is $L(\overline{c}, \mu)$ if and only if H is $L(c, \mu)$.*

Proof By the theorem, if H is $L(\overline{c}, \mu)$ then H has finite VC dimension, and therefore, by Theorem 3.15, H is learnable; that is, H is $L(c, \mu)$. The converse is plain since for any H, if H is $L(c, \mu)$ then H is certainly $L(\overline{c}, \mu)$. □

Therefore the class of hypothesis spaces learnable uniformly over distribution but not over target concept is empty. That is, if any particular concept in H is μ-learnable in H for all μ, with a value of m_0 independent of μ, then H is $L(c, \mu)$; that is, H is learnable.

6.3 Distribution-Dependent Learnability

The aim of this section is to show, first by considering the expectation of the index function and then by presenting a general theory of Ben-David *et al* [7], that, in contrast to the negative result of the previous section, allowing learnability to be distribution-dependent does indeed bring some new hypothesis spaces into consideration.

Distribution-dependent learnability and index functions

In chapter 3, in search of learnability results and sample-size bounds which were independent of distribution, we made use of the bound

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < C(r,k)\, \Pi_H(m) \binom{k}{\lceil r\epsilon k \rceil} \binom{m+k}{\lceil r\epsilon k \rceil}^{-1}$$

of Theorem 3.3. For distribution-dependent analysis, if Π_H is measurable (which is certainly the case if H is universally separable) we can use $\mathbf{E}(\Pi_H(\mathrm{x}))$ in place of $\Pi_H(m)$. Here, as earlier, $\mathbf{E}(\Pi_H(\mathrm{x}))$ denotes the expected value (with respect to μ^{m+k} and over X^{m+k}) of $\Pi_{m+k,H}$.

A function f is said to be *subexponential* if, for all $\epsilon > 0$, as x tends to infinity, $f(x)\exp(-\epsilon x)$ tends to zero. With this definition, we have the following.

Theorem 6.4 *Let μ be any probability measure on X. If $\mathbf{E}(\Pi_{n,H}(\mathrm{x}))$, the expected value of $\Pi_{n,H}(\mathrm{x})$ over X^n (with respect to μ^n) is a subexponential function of n, then any concept $c \in H$ is μ-learnable, with a sufficient sample size m_0 independent of c.*

Proof In view of the above discussion, proceeding as in Chapter 3 and modifying Corollary 3.9 in the obvious way, we obtain: For all $m \geq 8/\epsilon$,

$$\mu^m \{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < 2\,\mathbf{E}(\Pi_{2m,H}(\mathrm{x}))\, 2^{-\epsilon m/2}.$$

If

$$\mathbf{E}(\Pi_{2m,H}(\mathrm{x}))\; 2^{-\epsilon m/2} \to 0 \;\text{ as }\; m \to \infty,$$

for all $\epsilon > 0$, which is certainly the case if $\mathbf{E}(\Pi_{n,H}(\mathrm{x}))$ is a subexponential function of n, then the quantity on the right-hand side can be made less than any $\delta > 0$ by choosing $m \geq m_0$, say. Since the right-hand side of the inequality does not depend on c, m_0 can be chosen independently of c. Therefore any $c \in H$ is μ-learnable, with a sufficient sample-size uniform over all $c \in H$. □

As an application of this theorem, we give an example of a hypothesis space H of infinite VC dimension and a particular distribution μ on the input space, for which Theorem 6.4 implies the μ-learnability of any concept in H.

Let $\{B_n\}_{n\geq 1}$ be any sequence of disjoint sets such that $|B_i| = i$, $(i \geq 1)$. Take as input space the countably infinite set

$$X = \bigcup_{i=1}^{\infty} B_i,$$

and let the probability measure μ be defined on the σ-algebra of all subsets of X by

$$\mu(\{x\}) = \frac{1}{i}\frac{1}{2^i} \ (x \in B_i).$$

Let the hypothesis space H be the set of functions

$$H = \bigcup_{i=1}^{\infty} \{I_C : C \subseteq B_i\},$$

where $I_C : X \to \{0,1\}$ is the indicator (or characteristic) function of the subset C. Then:

Lemma 6.5 *The hypothesis space H defined above has infinite VC dimension.*

Proof For each positive integer m, the set B_m is shattered by H and therefore H has infinite VC dimension. □

Proposition 6.6 *Any concept $c \in H$ is μ-learnable, with a sufficient sample-size independent of c.*

Proof For $\mathrm{x} \in X^m$, let $I(\mathrm{x})$ be the set of entries of x. That is, $I(\mathrm{x}) = \{x_i : 1 \leq i \leq m\}$. Then it is not difficult to see that

$$\Pi_H(\mathrm{x}) = \sum 2^{|I(\mathrm{x})\cap B_i|},$$

where the sum is over all i such that $I(\mathrm{x}) \cap B_i \neq \emptyset$. Therefore,

$$I(\mathrm{x}) \subseteq S_k = \bigcup_{i=1}^{k} B_i \Longrightarrow \Pi_H(\mathrm{x}) \leq 2 + 2^2 + \ldots + 2^k < 2^{k+1}.$$

Further, $\Pi_H(\mathrm{x}) \leq 2^m$ for all $\mathrm{x} \in X^m$.

Let η_k be the probability that $I(\mathrm{x}) \subseteq S_k$; that is, $\eta_k = \mu^m(S_k^m)$. Then,

$$\eta_k = (\mu(S_k))^m = \left(1 - \frac{1}{2^k}\right)^m.$$

For any $0 < x < 1$,

$$(1-x)^m \geq 1 - mx.$$

Therefore, for $k \geq 2$,

$$\eta_k - \eta_{k-1} \leq 1 - \left(1 - \frac{1}{2^{k-1}}\right)^m \leq \frac{m}{2^{k-1}}.$$

Since the sets S_k^m cover X^m, we therefore have

$$\begin{aligned}
\mathbf{E}(\Pi_H(\mathrm{x})) &< 2\eta_1 + \sum_{k=2}^{m-1} (\eta_k - \eta_{k-1})\, 2^{k+1} + \sum_{k=m}^{\infty} (\eta_k - \eta_{k-1})\, 2^m \\
&\leq 1 + \sum_{k=2}^{m-1} \frac{m}{2^{k-1}} 2^{k+1} + \sum_{k=m}^{\infty} \frac{m}{2^{k-1}} 2^m \\
&= 1 + 4m(m-2) + 4m \\
&< 4m^2.
\end{aligned}$$

It follows that the expected value of $\Pi_H(\mathrm{x})$ over all $\mathrm{x} \in X^m$, with respect to the measure μ^m, is polynomial and therefore, by Theorem 6.4, any $c \in H$ is μ-learnable in H with a value of m_0 independent of c. □

Thus we see that it is possible to have every concept in a hypothesis space learnable with respect to a particular distribution and to have this learnability uniform over the concepts, even when the hypothesis space has infinite VC dimension. Since the hypothesis space has infinite VC dimension, Corollary 3.17 implies that it is not learnable; that is, H is not $L(c, \mu)$. We shall see in fact that the conclusion of Proposition 6.6 holds for *any* probability measure μ on X and not merely the particular one chosen above, so that H is $L(c, \overline{\mu})$. Note that the sample-size, the m_0 in the definition of $L(c, \overline{\mu})$, *must* depend on the distribution μ since H is not learnable.

Hypothesis spaces of $X\sigma$-finite dimension

We now present a theory due to Ben-David, Benedek and Mansour [7], who introduced the idea of $X\sigma$-finite dimensional hypothesis spaces. For a hypothesis space H defined over an input space X, the notation $H|Y$, for $Y \subseteq X$, shall denote the restriction of H to domain Y.

Definition 6.7 *A hypothesis space H over an input space X is said to have $X\sigma$-finite dimension if there is a countable family*

$$\{B_i\}_{i=1}^{\infty}$$

of subsets of X such that

$$\mathrm{VCdim}\,(H|B_i) < \infty$$

and

$$\bigcup_{i=1}^{\infty} B_i = X.$$

Here, $H|B_i$ denotes the restriction of H to domain B_i. □

Consider again the example of the previous subsection. The input space X is the disjoint union of sets B_i, where B_i has cardinality i, and the hypothesis space H is the collection

$$H = \bigcup_{i=1}^{\infty} \{I_C : C \subseteq B_i\}$$

of all indicator functions of the subsets of the sets B_i. Each of the sets B_i is shattered by H, and so the VC dimension of $H|B_i$ is equal to i. Thus X is the countable union of sets on which H has finite VC dimension; that is, H has $X\sigma$-finite dimension. Here, of course, $H|B_i$ has finite VC dimension since each B_i is finite. However, in general, the B_i of the definition need not be finite.

The following result is proved in [7]. The proof follows from the proof of a theorem in the next section and so we shall omit it here.

Theorem 6.8 *If hypothesis space H has $X\sigma$-finite dimension then H is $L(c, \overline{\mu})$.* □

In particular, this theorem shows:

Corollary 6.9 *If H is a hypothesis space over a countable set, then H is $L(c, \overline{\mu})$.*

Proof Suppose that H is defined over the countable set X. Then H certainly has $X\sigma$-finite dimension. For, take the B_i of the definition to be the singleton subsets of X. The VC dimension of H restricted to a singleton set is at most one, and X is the countable union of its singleton subsets. The result follows from Theorem 6.8. □

Corollary 6.9 provides a proof, mentioned earlier, of our claim that the result of Proposition 6.6 holds for any distribution μ, and it therefore shows that the class of hypothesis spaces which are $L(c, \overline{\mu})$ and not learnable is non-empty. Thus the notion of distribution-dependent learnability is not a vacuous one.

Corollary 6.9 shows that any hypothesis space defined over a countable input space has $X\sigma$-finite dimension. We now give an example of a hypothesis space H over a (necessarily) uncountable input space X such that H is not $X\sigma$-finite dimensional. Take X to be the closed interval $X = [0, 1]$, and let H be the space of all (characteristic functions of) subsets of X. Now, H shatters any subset of X and therefore, for any $Y \subseteq X$,

$$\mathrm{VCdim}\,(H|Y) = |Y|.$$

If X were the countable union

$$X = \bigcup_{i=1}^{\infty} B_i$$

of sets B_i such that H had finite VC dimension on B_i then, in particular, each B_i would be finite and X, as the countable union of finite sets would be countable. However, X is uncountable and we therefore deduce that H does not have $X\sigma$-finite VC dimension.

Polynomial Distribution-Dependent Learnability

Theorem 6.8 provides a positive distribution-dependent learnability result. However, it does not address the size of sample required for learnability to given degrees of accuracy and confidence. A closer analysis of the proof of this result in [7] shows that the resulting sufficient sample-size will not be polynomial in $1/\epsilon$ and $1/\delta$ for many distributions. The learnability results for spaces of finite VC dimension in chapters 3 and 5 provide sample-size bounds polynomial in these parameters, and this is desirable for the reasons mentioned in Chapter 1: if a learning algorithm is to run in time polynomial in $1/\epsilon$ and $1/\delta$, it must take as input a sample of size at most polynomial in these parameters. We therefore make the following definition:

Definition 6.10 *Let H be a hypothesis space over input space X and let $\mathcal{D}$ be a set of probability measures (distributions) defined on X. Then H is polynomially $L(c,\overline{\mu})$ with respect to $\mathcal{D}$ if H is $L(c,\overline{\mu})$ and if for any $c \in H$, for any $\epsilon, \delta \in (0,1)$ and for any μ in $\mathcal{D}$, there is an integer $m_0 = m_0(\epsilon, \delta, \mu)$, polynomial in $1/\epsilon$ and $1/\delta$, such that for all $m \geq m_0$,*

$$\mu^m\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\} < \delta.$$

□

To introduce the approach taken here, we prove the following.

Proposition 6.11 *H has $X\sigma$-finite dimension if and only if there exists an increasing sequence $\{S_k\}_{k=1}^{\infty}$ of subsets of X such that*

$$\mathrm{VCdim}\,(H|S_k) \leq k$$

and

$$\bigcup_{k=1}^{\infty} S_k = X.$$

Proof Suppose that H has $X\sigma$-finite dimension, and let the sets B_i be as in Definition 6.7. Let $x_0 \in B_1$ and set $B_0 = \{x_0\}$. For $k \geq 1$ let

$$S_k = \bigcup_{i=0}^{m(k)} B_i,$$

where

$$m(k) = \max\left\{ m : H\Big|\bigcup_{i=0}^{m} B_i \text{ has VC dimension} \leq k \right\}.$$

Given any $x \in X$, there is an m such that $x \in \bigcup_{i=0}^{m} B_i$. Let

$$\text{VCdim}\left(H\Big|\bigcup_{i=0}^{m} B_i \right) = k < \infty.$$

Then $m(k) \geq m$, so $x \in S_k$.

Conversely, if such sets S_i exist, take $B_i = S_i$. Then $\text{VCdim}(H|B_i)$ is finite, and $\bigcup_{i=1}^{\infty} B_i = X$. □

If H "nearly" has finite VC dimension, in some sense, we might hope to get polynomially bounded sample-sizes. Therefore, motivated by Proposition 6.11, we make the following definition.

Definition 6.12 *H has polynomial $X\sigma$-finite dimension with respect to probability measure μ if there exists an increasing sequence $\{S_k\}_{k=1}^{\infty}$ of subsets of X such that*

$$\text{VCdim}(H|S_k) \leq k,$$

$$\bigcup_{k=1}^{\infty} S_k = X$$

and

$$k(\alpha) = \min\{k : \mu(S_k) \geq 1 - \alpha\} \leq P\left(\frac{1}{\alpha}\right),$$

for some polynomial P.

If $\mathcal{D}$ is a set of probability measures defined on X, then *H has polynomial $X\sigma$-finite dimension with respect to $\mathcal{D}$* if H has polynomial $X\sigma$-finite dimension with respect to each μ in $\mathcal{D}$. □

Benedek and Itai [8] have gone some way towards investigating sample-sizes for distribution-dependent learnability, but only for the case of discrete distributions (that is, distributions nonzero on only countably many elements of the input space). With the definition of polynomial $X\sigma$-finite dimension, we can formalise some of the ideas they have considered and develop a theory for non-discrete as well as discrete distributions.

The following holds:

Theorem 6.13 *Let H be a hypothesis space over X, and $\mathcal{D}$ a class of probability measures defined on X. If H has polynomial $X\sigma$-finite dimension with respect to $\mathcal{D}$, then H is polynomially $L(c,\overline{\mu})$ with respect to $\mathcal{D}$.*

Proof Suppose that H has polynomial $X\sigma$-finite dimension with respect to the class $\mathcal{D}$, and let μ be a particular probability measure from $\mathcal{D}$. Suppose that $0 < \alpha < 1$ and $S \subseteq X$ is such that $\mu(S) \geq 1 - \alpha$. The probability (with respect to μ^m) that a sample of length $m = 2l$, chosen according to μ, has at least half of its members in S is at least

$$1 - \sum_{k=0}^{l} \binom{2l}{k} \alpha^{2l-k}(1-\alpha)^k.$$

Now,

$$\begin{aligned}\sum_{k=0}^{l} \binom{2l}{k} \alpha^{2l-k}(1-\alpha)^k &\leq \sum_{k=0}^{l} \binom{2l}{k} \alpha^{2l-k} \\ &\leq \alpha^{2l-l} \sum_{k=0}^{l} \binom{2l}{k} \\ &= \alpha^l 2^{2l-1}.\end{aligned}$$

Therefore, this probability is at least

$$1 - \alpha^l 2^{2l-1}.$$

If

$$\alpha = \min\left(\frac{1}{4}\delta^{-\epsilon/\log\epsilon}, \frac{\epsilon}{2}\right) \quad \text{and} \quad l \geq l_0 = \left\lceil \frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right) \right\rceil + 1,$$

then

$$l(\log_2 \alpha + 2) < \frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right) \left(\frac{\epsilon}{\log(\frac{1}{\epsilon})}\right) \log_2 \delta = \log_2 \delta.$$

This implies that the above probability is greater than $1 - \delta/2$. For,

$$\begin{aligned}
&1 - \alpha^l 2^{2l-1} > 1 - \frac{\delta}{2} \\
&\iff \alpha^l 2^{2l-1} < \frac{\delta}{2} \\
&\iff l \log_2 \alpha + 2l - 1 < \frac{\delta}{2} - 1 \\
&\iff l(\log_2 \alpha + 2) < \log_2 \delta.
\end{aligned}$$

Therefore, with probability at least $1-\delta/2$, a random sample of length $m \geq 2l_0$ has at least half of its members in $S = S_{k(\alpha)}$. Let

$$m_0 = m_0(\epsilon, \delta, \mu) = \left\lceil \frac{2}{\epsilon(1-\sqrt{\epsilon})} \left(2k(\alpha) \log\left(\frac{12}{\epsilon}\right) + \log\left(\frac{4}{\delta}\right)\right)\right\rceil.$$

Suppose that we choose $c \in H$ as the target concept. Since $H|S$ has VC dimension at most $k(\alpha)$, m_0 is twice a sufficient sample size for the learnability of $H|S$ with accuracy $\epsilon/2$ and confidence $1 - \delta/2$. Let $m \geq m_0$, and let $l = \lfloor m/2 \rfloor \geq l_0$. If $\mathrm{x} \in X^m$ is such that x has at least l of its entries from $S = S_{k(\alpha)}$, then we shall denote by x_S the unique vector of length l whose entries are precisely the first l entries of x from S, appearing in the same order as in x. Let μ_1 be the probability measure induced on S by μ. Thus, for any measurable subset A of X,

$$\mu_1(A \cap S) = \frac{\mu(A)}{\mu(S)}.$$

Observe that if $h \in H$ is such that $h \in H[\mathrm{x}]$ and $\mathrm{er}_\mu(h) > \epsilon$ then, since

$$\mu(S) \geq 1 - \alpha \geq 1 - \epsilon/2,$$

the function $h|S$ (h restricted to S) is such that

$$h|S \in (H|S)[\mathrm{x}_S]$$

and

$$\begin{aligned}\mathrm{er}_{\mu_1}(h|S) &= \frac{1}{\mu(S)}\mu\left(\{x \in S : h(x) \neq c(x)\}\right) \\ &= \frac{1}{\mu(S)}\mu\left(\{x \in X : h(x) \neq c(x)\} \cap S\right) \\ &> \frac{1}{\mu(S)}\left(\epsilon - \frac{\epsilon}{2}\right) \\ &> \frac{\epsilon}{2}.\end{aligned}$$

Therefore, denoting the number of entries of a vector x which lie in S by $s(\mathrm{x})$, we have

$$\begin{aligned}&\mu^m\left\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon\right\} \\ &= \mu^m\left\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon \text{ and } s(\mathrm{x}) \geq l\right\} \\ &\qquad + \mu^m\left\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon \text{ and } s(\mathrm{x}) < l\right\}.\end{aligned}$$

The second measure here is at most $\delta/2$ since with probability at least $1-\delta/2$, $s(\mathrm{x})$ is at least l. Further,

$$\begin{aligned}&\mu^m\left\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon \text{ and } s(\mathrm{x}) \geq l\right\} \\ &= \mu^m\left\{\mathrm{x} \in X^m : \mathrm{haz}_\mu(H[\mathrm{x}]) > \epsilon \middle| s(\mathrm{x}) \geq l\right\} \mu^m\left\{\mathrm{x} \in X^m : s(\mathrm{x}) \geq l\right\} \\ &\leq \mu^m\left\{\mathrm{x} \in X^m : \exists h \in H[\mathrm{x}] \text{ with } \mathrm{er}_\mu(h) > \epsilon \middle| s(\mathrm{x}) \geq l\right\} \\ &\leq \mu^m\left\{\mathrm{x} \in X^m : \exists f \in (H|S)[\mathrm{x}_S] \text{ with } \mathrm{er}_{\mu_1}(f) > \epsilon/2\right\},\end{aligned}$$

where, for any events A and B, $\mu^m(A|B)$ is the conditional probability (with respect to μ^m) of A given B.

Now, if $s(\mathrm{x}) \geq l$ and x is μ-randomly chosen, then x_S is a μ_1-randomly chosen sample of length l. Therefore this last measure is at most $\delta/2$, since l is a sufficient sample-size for the learnability of $H|S$ to accuracy $\epsilon/2$ with confidence $\delta/2$.

Thus, H is $L(c, \overline{\mu})$. Now, since H has polynomial $X\sigma$-finite dimension with respect to μ, $k(\alpha)$ is polynomially bounded in $1/\epsilon$ and $1/\delta$ and hence so also is $m_0(\epsilon, \delta, \mu)$. Since this holds for all μ in $\mathcal{D}$, H is polynomially $L(c, \overline{\mu})$ with respect to $\mathcal{D}$. □

The proof of this theorem also provides a proof of Theorem 6.8. This can be achieved by using the alternative characterization given in Proposition 6.11 of

spaces of $X\sigma$-finite dimension, and then repeating the proof without the last paragraph.

It is also perhaps worth noting that if, in the definition of polynomial $X\sigma$-finite dimension, the VC dimension of H restricted to S_k is allowed to be at most $q(k)$ for some polynomial q, rather than just k, the same result follows. This follows directly from the above proof; $q(k(\alpha))$ would replace $k(\alpha)$ in the sample-size bound.

To illustrate the idea of polynomial $X\sigma$-finite dimension, we now give an example of a hypothesis space H, together with a distribution μ such that H has polynomial $X\sigma$-finite dimension with respect to μ. It is the same example as earlier. The input space X is the disjoint union of the sets B_i, where B_i is of cardinality i, and the hypothesis space is the collection of all (characteristic functions of) the subsets of each B_i. Note that H certainly has $X\sigma$-finite dimension since X is countable. Define the probability measure μ as before;

$$\mu(\{x\}) = \frac{1}{i}\frac{1}{2^i} \ (x \in B_i).$$

For each k, let

$$S_k = \bigcup_{i=1}^{k} B_i.$$

Then $\{S_k\}_{k=1}^{\infty}$ is an increasing sequence of subsets of X such that

$$\bigcup_{k=1}^{\infty} S_k = X.$$

Further, if a subset of S_k is shattered, that subset must lie entirely within one of the $B_i (1 \leq i \leq k)$ and hence

$$\begin{aligned} \mathrm{VCdim}(H|S_k) &= \max\{\mathrm{VCdim}(H|B_j) : j \leq k\} \\ &= \mathrm{VCdim}(H|B_k) = k. \end{aligned}$$

Now,

$$\mu(S_k) = 1 - \frac{1}{2^k},$$

from which it follows that, for $0 < \alpha < 1$,

$$k(\alpha) = \left\lceil \log_2 \left(\frac{1}{\alpha} \right) \right\rceil,$$

which is certainly bounded polynomially in $1/\alpha$. Thus H has polynomial $X\sigma$-finite dimension with respect to μ.

We have given an example of a hypothesis space H over an input space X such that H has infinite VC dimension and $X\sigma$-finite dimension. We have also given an example of a hypothesis space (over an input space X) which does not have $X\sigma$-finite dimension. Above, we provided an example of a hypothesis space H (over input space X), together with a probability measure μ on the input space, such that H has polynomial $X\sigma$-finite dimension with respect to μ. It remains to give an example of a hypothesis space H over an input space X, together with a probability distribution μ on X, such that H has $X\sigma$-finite dimension but does *not* have polynomial $X\sigma$-finite dimension with respect to μ.

Let X be the set of all natural numbers and H the set of all subsets of X. The input space is countable, and therefore H has $X\sigma$-finite dimension. Define the probability measure μ on X by

$$\mu(\{x\}) = \frac{1}{\log x} - \frac{1}{\log(x+1)} \ (x > 1), \quad \mu(\{1\}) = 1 - \frac{1}{\log 2}.$$

Suppose that the sequence of sets $\{S_k\}_{k=1}^{\infty}$ is such that

$$X = \bigcup_{k=1}^{\infty} S_k \text{ and } \mathrm{VCdim}\,(H|S_k) \leq k.$$

Every subset of X is shattered by H, so that

$$\mathrm{VCdim}\,(H|S_k) = |S_k|\,.$$

But H restricted to S_k is supposed to have VC dimension at most k. Therefore, for each integer k, S_k has cardinality at most k. It follows that

$$\mu\,(S_k) \leq \mu\,(\{1, 2, \ldots, k\}) = 1 - \frac{1}{\log(k+1)}.$$

From this, we obtain

$$k(\alpha) \geq \left\lfloor \exp\left(\frac{1}{\alpha}\right) \right\rfloor,$$

which is not bounded by any polynomial in $1/\alpha$. It follows that H does not have polynomial $X\sigma$-finite dimension with respect to μ.

Given the sequence $\{S_k\}_{k=1}^{\infty}$, the crucial quantity in the above analysis is the function f defined by

$$f(k) = 1 - \mu(S_k).$$

If $f(k)$ tends to 0 "fast enough" as k tends to infinity, then the theorem guarantees a sample-size polynomial in $1/\epsilon$ and $1/\delta$. Formally, if

$$f(k) = O\left(\frac{1}{k^c}\right)$$

for some constant $c > 0$, then

$$k(\alpha) = O\left(\left(\frac{1}{\alpha}\right)^{1/c}\right),$$

polynomial in $1/\alpha$. But if

$$f(k) = \Omega\left(\frac{1}{(\log k)^c}\right)$$

then

$$k(\alpha) = \Omega\left(\exp\left(\left(\frac{1}{\alpha}\right)^{1/c}\right)\right),$$

and the theorem does not guarantee a polynomial sample size.

Chapter 7
Learning Formal Concepts

7.1 Introduction

Formal Concepts were introduced by Wille [37] to capture the philosophical ideas of intent and extent in a lattice-theoretic framework. In this chapter, we discuss formal concept analysis and show that learnability results can be applied to give results on the learnability of the space of formal concept extents. We investigate the relationship between the VC dimension of these spaces and the structure of the underlying context, showing that in certain cases we can bound the VC dimension.

We also show that formal concepts can be regarded as a generalization of monomials, and that monomial learning algorithms [33, 16] can be adapted to yield efficient algorithms for learning concepts in finite contexts.

It is intended that this chapter illustrate much of the theory developed in the preceeding chapters.

7.2 Formal Concept Analysis

Contexts and formal concepts

Let X and A be (possibly infinite) sets, whose members we will call *objects* and *attributes* respectively. Let I be a subset of $X \times A$. Wille [37] calls the triple (X, A, I) a *context*. Thus a context can be regarded as an incidence structure defined on sets X and A with incidence given by I. We shall write xIa to mean $(x, a) \in I$. This can be thought of as meaning "object x has attribute a".

Given a context (X, A, I), we have *incidence operators* I_X and I_A, defined on the power sets of X, A respectively.

Definition 7.1 *For a context (X, A, I), $I_X : 2^X \to 2^A$ is defined for $C \subseteq X$ by*

$$I_X(C) = \{a \in A : cIa \ \ \forall c \in C\}$$

and $I_A : 2^A \to 2^X$ is defined for $D \subseteq A$ by

$$I_A(D) = \{x \in X : xId \ \ \forall d \in D\}.$$

□

Thus, for $C \subseteq X$, $I_X(C)$ is the largest set of attributes shared by the members of C and for $D \subseteq A$, $I_A(D)$ is the largest set of objects sharing the attributes of D. We shall use the symbol I to denote each of I_X and I_A, as it will usually be clear which is meant.

The ordered pair (C, D) is a said to be a formal concept if the objects of C all share the attributes in D and no others, and the objects sharing the attributes of D are precisely the objects in C. More formally,

Definition 7.2 *A pair (C, D) with $C \subseteq X$ and $D \subseteq A$ is a formal concept belonging to the context (X, A, I) if $I(C) = D$ and $I(D) = C$.* □

We regard (X, ϕ) and (ϕ, A) as formal concepts.

Using terms borrowed from philosophy, C is called the *extent* and D the *intent* of the formal concept (C, D). If (C, D) is a formal concept, then the pair (C, D) has a certain maximality with respect to the context; C is maximal among all subsets C' of X for which every member of C' is incident with every attribute in D and, dually, D is maximal among all subsets D' of A for which every member of D' is incident with every object in C. Notice that the extent of a formal concept uniquely determines the intent and, dually, the intent uniquely determines the extent. Therefore either the extent or the intent serves to uniquely define a formal concept. We denote the set of formal concepts belonging to (X, A, I) by $FC(X, A, I)$.

When X and A are countable, the context can be represented by a $(0, 1)$-array, which we call the *context table*. The rows are indexed by X and the columns by A, with a one in position (x, a) if and only if xIa. Thus, the context table is the incidence array of the context. Let us call a sub-array of the context table a *block* if every entry of the sub-array is a one. A formal concept (C, D) gives rise to a block; namely the sub-array formed by the rows indexed by C and the columns indexed by D. The definition of formal concept implies that this block is maximal. That is, it is not a sub-array of any strictly larger block. Conversely, any maximal block gives rise to a unique formal concept in the obvious way.

Example

To illustrate these definitions, consider the following finite context table, where the rows have been labelled with the objects

$$X = \{x_1, x_2, \ldots, x_7\}$$

and the columns with the attributes

$$A = \{a_1, a_2, \ldots, a_6\}.$$

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{array} \\ \left(\begin{array}{cccccc} 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{array} \right) \end{array}$$

Then

$$(\{x_1, x_3, x_4\}, \{a_2, a_3, a_5\})$$

is a formal concept belonging to the context, but neither of

$$(\{x_1, x_3\}, \{a_2, a_3, a_5\}),$$

$$(\{x_1, x_3, x_4\}, \{a_2, a_3\})$$

is a formal concept belonging to this context.

Properties of the incidence operators

We shall require some elementary results from [37] involving the I operators.

Proposition 7.3 *If $C_1 \subseteq C_2 \subseteq X$ then $I(C_1) \supseteq I(C_2)$. Similarly, if $D_1 \subseteq D_2 \subseteq A$ then $I(D_1) \supseteq I(D_2)$.*

Proof Suppose $C_1 \subseteq C_2 \subseteq X$. If $a \in I(C_2)$ then cIa for all $c \in C_2$. Therefore cIa for all $c \in C_1$. Thus $I(C_2) \supseteq I(C_1)$. The proof of the second part is similar. □

Proposition 7.4 *For $C \subseteq X$ and $D \subseteq A$, $C \subseteq I^2(C)$ and $D \subseteq I^2(D)$.*

Proof This is clear. For, each object in C has each of the attributes in $I(C)$ (by definition of $I(C)$), and therefore $C \subseteq I^2(C)$. The other part is dual to this. □

Proposition 7.5 *For any $C \subseteq X$ and $D \subseteq A$, $I^3(C) = I(C)$ and $I^3(D) = I(D)$.*

Proof We prove the first half of the result, the second half being dual to this. By the preceeding proposition, $C \subseteq I^2(C)$ and so, since I reverses inclusion,

$$I^3(C) = I(I^2(C)) \subseteq I(C).$$

Also, for any $D \subseteq A$, $D \subseteq I^2(D)$. Taking $D = I(C)$ gives $I(C) \subseteq I^3(C)$. Therefore $I^3(C) = I(C)$. □

We shall denote the set of all extents of formal concept belonging to the context (X, A, I) by $\mathcal{E}(X, A, I)$. The formal concept extents can be characterized as those subsets C of X for which $I^2(C) = C$, because $FC(X, A, I)$ consists precisely of the pairs $(C, I(C))$ where $I(I(C)) = C$. We shall often identify concept extents with their characteristic functions in what follows.

Formal concepts and monomials

The extents of the formal concepts in $FC(X, A, I)$ are precisely the subsets of X of the form $I(D)$ for some subset D of A. For, given any $D \subseteq A$, we certainly have $I(D) \subseteq X$, and $I^2(D) = I(I(D)) \subseteq A$. Further, $I\left(I(D)\right) = I^2(D)$ (by definition) and

$$I\left(I^2(D)\right) = I^3(D) = I(D),$$

by Proposition 7.5, so that $\left(I(D), I^2(D)\right)$ is a formal concept belonging to the context. Therefore, for any subset D of A, $I(D)$ is a formal concept extent. Conversely, if $C \subseteq X$ is the extent of the formal concept (C, D) then, by definition, $C = I(D)$.

This result enables the idea of formal concepts to be related to that of monomials. For $x \in X$ let $\hat{x}$ denote the element of $\{0,1\}^A$ (that is, the function from A to $\{0,1\}$) such that

$$\hat{x}(a) = 1 \iff xIa.$$

When A is countably infinite, we regard $\hat{x} = (x_a)_{a \in A}$ as a $(0,1)$-sequence, and when A is finite, we regard it as a Boolean vector. Let (C, D) be a formal concept belonging to the context. Then

$$\begin{aligned} x \in C &\iff \{x\} \subseteq C \\ &\iff I(\{x\}) \supseteq I(C) = D \\ &\iff \hat{x}(d) = 1 \ \ \forall d \in D. \end{aligned}$$

When X is countable we may regard the set $\{x_a : a \in A\}$ as a countable set of Boolean variables. For D finite,

$$x \in C \iff f^D(\hat{x}) = 1$$

where

$$f^D(\hat{x}) = \prod_{d \in D} x_d,$$

is the monotone monomial formed as the conjunction of the variables x_d $(d \in D)$. Therefore C may be identified with a subset of the set of positive examples of a monotone monomial of finite length over a countable set of Boolean variables. This observation will prove useful in a later section, where we describe an efficient learning algorithm for learning formal concept extents in a given finite context.

Consider again the earlier context. With the above notation, $\hat{x}_1$ is the Boolean vector which is the first row of the context table. That is,

$$\hat{x}_1 = (0,1,1,0,1,0)\,.$$

We similarly define $\hat{x}_2, \ldots, \hat{x}_7$. Now,

$$(\{x_1, x_3, x_4\}, \{a_2, a_3, a_5\})$$

is a formal concept belonging to this context. The extent C of this concept is

$$C = \{x_1, x_3, x_4\}\,.$$

Let $(\hat{x}_i)_j$ denote entry j of $\hat{x}_i$, for $1 \leq j \leq 6$. Then C can be expressed as the set of x_i such that

$$(\hat{x}_i)_2(\hat{x}_i)_3(\hat{x}_i)_5 = 1.$$

That is,

$$C = \left\{x_i : f^{2,3,5}(\hat{x}_i) = 1\right\},$$

where, for $\mathrm{y} = (y_1, y_2, \ldots, y_6) \in \{0,1\}^6$,

$$f^{2,3,5}(y_1, y_2, \ldots, y_6) = y_2 y_3 y_5.$$

Then $f = f^{2,3,5}$ is a monotone monomial and C can be identified with a subset of the set of positive examples of f.

7.3 The Dimension of a Context

Introduction

We have seen in Chapter 3 that a hypothesis space is learnable if and only if it has finite VC dimension, and that if it has finite VC dimension, then the sample-size for learnability can be bounded in terms of the VC dimension. In this section, we discuss the VC dimension of the set of extents of formal concepts belonging to a given context, relating this to the structure of the context. We show that in contexts with certain boundedness properties the VC dimension of the space of formal concept extents can be bounded and therefore that learnability results can be obtained.

For ease of notation, we make the following definition.

Definition 7.6 *The dimension of the context* (X, A, I) *is defined to be the VC dimension of* $\mathcal{E}(X, A, I)$. □

Thus, the dimension of the context is a measure of the expressive power of the context or of the degree of classification of the objects provided by the context.

Boundedness properties and dimension

We shall find a sufficient condition for a context to have finite VC dimension (and thus for $\mathcal{E}(X, A, I)$ to be learnable).

Definition 7.7 *The context (X, A, I) is uniformly K-bounded (where K is a positive integer) if each object has at most K attributes. That is,*

$$\forall x \in X, \quad |I(\{x\})| \leq K.$$

□

Thus, in a countable uniformly K-bounded context, each row of the context table has at most K ones, and any formal concept extent is a subset of the set of positive examples of a monotone monomial which is the product of at most K literals. We may weaken this boundedness condition slightly, allowing a "small" number of objects to have a large or unbounded number of attributes.

Definition 7.8 *The context (X, A, I) is almost uniformly K-bounded if there are at most K objects which have at least $K + 1$ attributes. That is,*

$$|\{x \in X : |I(x)| > K\}| \leq K.$$

□

Thus, if (X, A, I) is almost uniformly bounded then there are less than $K + 1$ objects having more than K attributes. Thus, (in a countable context) there can be no square blocks of size $(K + 1) \times (K + 1)$ in the context table. We say that a context is almost uniformly bounded if it is almost uniformly K-bounded for some positive integer K. In an almost uniformly bounded context, it is possible (in the case of infinite A) for some of the objects to have an infinite number of attributes, so that in this case a formal concept extent does not necessarily correspond to a finite monotone monomial as above. However, in such contexts the space of concept extents has finite VC dimension, as the following result shows.

Theorem 7.9 *If a context is almost uniformly K-bounded then the context has dimension at most $K+1$.*

Proof Let x be any member of X such that $|I(\{x\})| \leq s$. We show that x can belong to at most 2^s concept extents. If $C \in \mathcal{E}(X, A, I)$, so that $(C, D) \in FC(X, A, I)$ for some $D \subseteq A$, then

$$x \in C \Longrightarrow I(\{x\}) \supseteq I(C) = D.$$

There can be at most 2^s such subsets D of A, this being the number of subsets of a set of cardinality s. Therefore, since C is determined by D, there can be at most 2^s such extents C.

Let $H = \mathcal{E}(X, A, I)$ and suppose that there is a subset

$$S = \{x_1, x_2, \ldots, x_{K+2}\}$$

of cardinality $K+2$ which is shattered by H. Then the collection

$$\{C \cap S : C \in \mathcal{E}(X, A, I)\}$$

consists of all subsets of S. This implies that each x_i belongs to at least 2^{K+1} formal concept extents. The context is almost uniformly K-bounded, so there are at most K of the x_i with at least $K+1$ attributes. Therefore at least one of the x_i has strictly less than $K+1$ attributes, and this x_i can belong to at most 2^K formal concept extents, contradicting the above. Therefore $\text{VCdim}(\mathcal{E}(X, A, I)) \leq K+1$. □

This theorem shows that $\mathcal{E}(X, A, I)$ is learnable if (X, A, I) is almost uniformly bounded. We now define another type of "finiteness" that a context may have.

Definition 7.10 *The context (X, A, I) is locally finite if each object has at most a finite number of attributes. That is,*

$$\forall x \in X, \quad I(\{x\}) \text{ is finite.}$$

Thus, in a countable locally finite context, each row of the context table has a finite number of ones. However, local finiteness is not a strong enough restriction to guarantee finite dimension of the context, as the following result of D. Cohen shows.

Theorem 7.11 *There is a locally finite context of infinite dimension.*

Proof The construction essentially concatenates contexts of each finite dimension. Let X be the disjoint union,

$$X = \bigcup_{i=1}^{\infty} (X_i \times \{i\}),$$

where $X_i = \{1, ..., i\}$ and let

$$A = \bigcup_{i=1}^{\infty} (A_i \times \{i\}),$$

where $A_i = \{0, 1, ..., i\}$. I is defined to be

$$I = \bigcup_{i=1}^{\infty} \{((x, i), (a, i)) : 1 \leq x \leq i, 0 \leq a \leq i, x \neq a\}.$$

Then (X, A, I) is certainly locally finite, but not almost uniformly bounded. It is easy to show that, for each k, the *subcontext* with object set $(X_k \times \{k\})$, attribute set $(A_k \times \{k\})$ and incidence induced by I (that is, the restriction of I), has dimension k. The dimension of the context is certainly at least as large as the dimension of any subcontext, and hence must be infinite. □

7.4 Non-Uniform Learnability of Formal Concepts

When the context is of infinite dimension and therefore not learnable, it is natural to consider non-uniform learnability. By the results of Chapter 6, we can not hope for learnability which is uniform over distribution. However, as shown there, we can guarantee distribution-dependent learnability in many cases. We illustrate this by discussing the context introduced in the proof of Theorem 7.11. By Theorem 6.9, since X is countable, $\mathcal{E}(X, A, I) \in L(c, \overline{\mu})$. However, we should like to consider polynomial distribution-dependent learnability. Using the notation introduced there, suppose that the probability measure μ on X is such that $(X_i \times \{i\})$ has measure p_i. Let $H = \mathcal{E}(X, A, I)$ and let $(X_i \times \{i\})$ be denoted B_i. Then for any i,

$$\text{VCdim}\,(H|B_i) = i.$$

Now, let

$$S_k = \bigcup_{i=1}^{k} B_i.$$

Then the sequence $\{S_k\}$ of sets is increasing, and

$$\text{VCdim}\,(H|S_k) = k.$$

By Theorem 6.13, if

$$k(\alpha) = \min\left\{k : \sum_{i=1}^{k} p_i \geq 1 - \alpha\right\} \leq P\left(\frac{1}{\alpha}\right)$$

for some polynomial P then $\mathcal{E}(X, A, I)$ is polynomially $L(c, \overline{\mu})$ with respect to μ.

For example, suppose that $p_i = 1/2^i$. Then

$$\mu(S_k) = \sum_{i=1}^{k} \frac{1}{2^i} = 1 - \frac{1}{2^k},$$

and therefore

$$k(\alpha) = \left\lceil \log_2\left(\frac{1}{\alpha}\right)\right\rceil,$$

which is certainly bounded polynomially in $1/\alpha$.

7.5 Learning Formal Concepts in a Finite Context

The context (X, A, I) is said to be finite if X and A are finite sets. Throughout this section, the context is assumed to be a fixed finite context (X, A, I) and we shall often identify a subset of X with its characteristic function. As mentioned above, formal concepts can be related to monomials. In this section we show how monomial learning algorithms can be modified in a straightforward way to obtain efficient algorithms for learning the formal concept extents belonging to the context. As described earlier, if $A = \{a_1, a_2, \ldots, a_t\}$, any particular formal concept extent can be identified with a subset of the set of positive examples of a monotone monomial over the $t = |A|$ boolean variables $\{x_{a_1}, ..., x_{a_t}\}$. Explicitly, we identify the concept extent $I(D)$ with the set of positive examples in X of the monotone monomial f^D, the conjunction of the variables x_d for d in D. The context is finite and therefore $\mathcal{E}(X, A, I)$ is finite and, consequently, (potentially) learnable. By modifying any efficient monomial-learning algorithm so that it returns a consistent concept extent, we can produce an efficient algorithm that learns $\mathcal{E}(X, A, I)$. To learn a particular member C of $\mathcal{E}(X, A, I)$, the algorithm $\mathcal{A}_I$, which "knows" the context (X, A, I), learns, using the sample x, a subset D of A that approximates $I(C)$. $\mathcal{A}_I$ then returns $I(D)$ as the approximating concept extent; that is, $\mathcal{A}_I(\mathrm{x}) = I(D)$.

Using the standard and simplest monomial learning algorithm [33], a suitable algorithm $\mathcal{A}_I$ is:

begin
$D := \{a_1, \ldots, a_t\}$
for each positive example x presented **do**
    begin
    for $i = 1$ to t **do**
        if $(x, a_i) \notin I$ **then** $D := D \setminus \{a_i\}$
    end
$\mathcal{A}_I(\mathrm{x}) := I(D)$
end

The following elementary observation is critical.

Lemma 7.12 *The algorithm $\mathcal{A}_I$ when given as input the sample* x *produces a member of $\mathcal{E}(X, A, I)$ consistent with the target extent on* x. *That is, $\mathcal{A}$ is a consistent algorithm.*

Proof By the earlier discussion, $I(D) \in \mathcal{E}(X, A, I)$. At each stage of the algorithm, the current D contains $I(C)$, the true intent. Therefore $I(D)$ is contained in $I^2(C) = C$, the true extent. So $I(D)$ correctly classifies the negative examples of the sample. (Indeed, it correctly classifies all negative examples of C). Let x be a positive example from the sample. At each stage after x has been presented to the algorithm, the current D is contained in $I(\{x\})$. Therefore,

$$\{x\} \subseteq I^2(\{x\}) \subseteq I(D).$$

That is, x is correctly classified by $I(D)$. □

Therefore, provided the number of examples input to this algorithm is at least $m_0(\epsilon, \delta)$, where this is the sufficient sample-size given in Theorem 3.15, the algorithm produces a concept extent which, with probability at least $1 - \delta$, has error less than ϵ with respect to the target extent. Notice that $\mathcal{E}(X, A, I)$ has VC dimension at most $n = |X|$, ~~since there are at most 2^n possible formal concept extents.~~ Therefore, by Theorem 3.15, we may take m_0 to be polynomially bounded in $1/\epsilon, 1/\delta$ and n. Since $I(D)$ can be computed from D in at most $|D|n$ operations, the worst-case time complexity of $\mathcal{A}_I$ on an input of m_0 examples is $O(|D|n + tm_0)$. This establishes the polynomial learnability of $\mathcal{E}(X, A, I)$. Notice that the worst-case time complexity also depends polynomially on the "size", $|X||A| = nt$ of the context.

We end this section by remarking that when $I(C)$ is small in size relative to A, it may be more efficient to use the algorithm developed by Haussler [16]. This method, which makes use of both the positive and the negative examples in the sample, is based upon a heuristic for the set-cover problem.

Chapter 8
The Learnability of Functions

8.1 Introduction

Instead of considering just $\{0,1\}$-valued functions, we should now like to consider functions taking values in some arbitrary set Y. We consider here only countable Y. One reason for this is that to apply the standard learnability framework and define a hypothesis to be in error on an input if its value on that input is not the same as the value of the target function on that input seems extremely restrictive when Y is, for example, the set of real numbers or some real interval. If we hope to achieve such exact correctness, we should perhaps deal with discrete and not continuous output spaces Y. Another reason is that when Y is countable, any measurable function c from X to Y, together with a probability measure μ on X, can be represented as a probability distribution on an appropriate σ-algebra over the product space $X \times Y$. This representation is not so explicit if Y is uncountable. The same sorts of upper bounds on sufficient sample size for learnability can be obtained as for the case of Boolean-valued functions, again in terms of a parameter that quantifies in some sense the "expressive power" of the space of functions. We use a definition of Haussler [17] for this parameter, which we continue to call the VC dimension. The aim of this short chapter is to provide a framework in which to apply learnability theory, in the next chapter, to particular types of artificial neural network.

8.2 Learnability Results for Function Spaces

A VC dimension for function spaces

We now discuss various approaches to finding a generalized definition of VC dimension which, in some sense, quantifies the expressive power of a set of functions from an input space X to an output space Y. For consistency, we want the generalized dimension to reduce to the straightforward definition of VC dimension when the range space has only two elements. Various definitions have been proposed.

If Y is finite, one possible definition which extends the earlier theory is as follows: Suppose that H is a set of functions from a set X to a finite set Y and define, for each $\mathbf{x} = (x_1, \ldots, x_m)$, the function

$$\mathbf{x}^* : H \to Y^m$$

by

$$\mathbf{x}^*(h) = (h(x_1), \ldots, h(x_m)).$$

Let

$$\Delta_H(m) = \max\{|\mathbf{x}^*(H)| : \mathbf{x} \in X^m\} \leq |Y|^m,$$

and define the VC dimension of H to be the largest integer d such that $\Pi_H(d) = |Y|^d$ (If such an integer exists; if not, define the dimension to be infinite). However, the condition $\Delta_H(m) = |Y|^m$ is a very demanding condition to be met and it seems that a space may have quite a large degree of expressibility, yet have a low VC dimension according to this definition.

Using definitions of Natarajan [25] we can give another possible definition of a VC dimension for function spaces with range in some arbitrary set Y. Suppose that H is a set of functions from a set X to a set Y. Natarajan defines a subset S of X to be "shattered" by H if there are two functions $f, g \in H$ such that

$$f(s) \neq g(s) \quad \forall s \in S,$$

and if for any $T \subseteq S$, there is a function $e \in H$ such that

$$e(s) = f(s) \ (s \in T), \quad e(s) = g(s) \ (s \in S \setminus T).$$

Thus the functions f and g separate the set S. We could then define the VC dimension of H to be the size of the largest subset of X shattered by H. However, as with the previous definition, this measure seems too small; the separation requirement is a very stringent one.

We therefore adopt a definition of Haussler [17], which involves a weaker separation condition. For any $h \in H$, the graph $\mathcal{G}(h)$ of h is the subset

$$\mathcal{G}(h) = \{(x, h(x)) : x \in X\}$$

of the product space $X \times Y$, and the *graph space* of H is the set of all graphs of functions in H,

$$\mathcal{G}(H) = \{\mathcal{G}(h) : h \in H\}.$$

We now define the VC dimension of the space H of functions from X to Y to be the VC dimension of the graph space $\mathcal{G}(H)$. This is (infinity, or) the size of the largest subset of $X \times Y$ which is shattered by $\mathcal{G}(H)$. Now

$$S = \{(x_1, a_1), (x_2, a_2), \ldots, (x_m, a_m)\} \subseteq X \times Y$$

is shattered by $\mathcal{G}(H)$ if and only if for any $T \subseteq \{1, 2, \ldots, m\}$, there is $\mathcal{G}(h_T) \in \mathcal{G}(H)$ such that for $1 \leq i \leq m$,

$$i \in T \Rightarrow (x_i, a_i) \in \mathcal{G}(h_T),$$

and

$$i \in S \setminus T \Rightarrow (x_i, a_i) \notin \mathcal{G}(h_T).$$

Thus, S is shattered by $\mathcal{G}(H)$ if and only if for each

$$\mathbf{b} = (b_1, b_2, \ldots, b_m) \in \{0, 1\}^m,$$

there is $h_{\mathbf{b}} \in H$ such that

$$h_{\mathbf{b}}(x_i) = a_i \iff b_i = 1.$$

The VC dimension of H is the cardinality of the largest subset of $X \times Y$ shattered by $\mathcal{G}(H)$.

An alternative description can be given. For $\mathbf{y} = (y_1, \ldots, y_m) \in Y^m$, let the function

$$I_{\mathbf{y}} : Y^m \to \{0,1\}^m$$

be defined by

$$I_{\mathbf{y}}((z_1, \ldots, z_m)) = (a_1, \ldots, a_m), \quad \text{where} \quad a_i = 1 \iff y_i = z_i.$$

As earlier, for $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$ and $h \in H$, define

$$\mathbf{x}^* : H \to Y^m$$

by

$$\mathbf{x}^*(h) = (h(x_1), h(x_2), \ldots, h(x_m)).$$

For each $\mathbf{y} \in Y^m$, the composition

$$(I_{\mathbf{y}} \circ \mathbf{x}^*) : H \to \{0,1\}^m$$

is a mapping from H to the finite set $\{0,1\}^m$. We then define, for $\mathbf{x} \in X^m$,

$$\Pi_{m,H}(\mathbf{x}) = \max \{|(I_{\mathbf{y}} \circ \mathbf{x}^*)(H)| : \mathbf{y} \in Y^m\}$$

to be the maximum, as $\mathbf{y}$ ranges over Y^m, of $|I_{\mathbf{y}} \circ \mathbf{x}^*(H)|$, the cardinality of the image of H under $(I_{\mathbf{y}} \circ \mathbf{x}^*)$. Such a quantity is well-defined, since the range of $(I_{\mathbf{y}} \circ \mathbf{x}^*)$ is finite. As in Chapter 2, we can, in the obvious way, define

$$\Pi_H : \bigcup_{m=1}^{\infty} X^m \to \mathbf{N}.$$

Further, we let

$$\Pi_H(m) = \max \{\Pi_H(\mathbf{x}) : \mathbf{x} \in X^m\} = \sup \Pi_{m,H}$$

be the maximum of $\Pi_H(\mathbf{x})$ over all $\mathbf{x} \in X^m$. Now, $\Pi_H(m) = 2^m$ if and only if for some $\mathbf{x} \in X^m$ and $\mathbf{y} \in Y^m$, $I_{\mathbf{y}} \circ \mathbf{x}^*(H)$ is the whole of $\{0,1\}^m$. This holds if and only if for any $\mathbf{b} \in \{0,1\}^m$, there is $h_{\mathbf{b}} \in \{0,1\}^m$ such that

$$h_{\mathbf{b}}(x_i) = y_i \iff b_i = 1.$$

Thus $\Pi_H(m) = 2^m$ if and only if $\mathcal{G}(H)$ shatters

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}.$$

Hence the VC dimension of H (is either infinite, or) is the largest integer d such that $\Pi_H(d) = 2^d$.

An observation which will prove useful later is that if Y is finite, then

$$\begin{aligned} \Pi_H(\mathbf{x}) &= \max\{|(I_{\mathbf{y}} \circ \mathbf{x}^*)(H)| : \mathbf{y} \in Y^m\} \\ &\leq |\mathbf{x}^*(H)| \\ &\leq \Delta_H(m), \end{aligned}$$

where $\Delta_H(m)$ is the maximum over all $\mathbf{x} \in X^m$ of $|\mathbf{x}^*(H)|$.

It is easy to see that if $Y = \{0,1\}$, this notion of VC dimension coincides with the standard one.

Stochastic and deterministic concepts with arbitrary range

As in Chapter 5, we can consider probability distributions on the set $X \times Y$ rather than functions from X to Y with underlying probability distributions on X. This allows us to discuss *stochastic concepts with range* Y, defined as probability measures on an appropriate product σ-algebra. When Y is countable, every pair (c, μ) where $c \in H$ and μ is a probability measure on X can be realised by a probability measure $\nu = \nu(c, \mu)$ on the product σ-algebra $\Phi = \Sigma \times 2^Y$. To see this, note that if $Y = \{y_n\}_{n=1}^{\infty}$ is an enumeration of Y then the product σ-algebra $\Phi = \Sigma \times 2^Y$ consists precisely of the sets of the form

$$\bigcup_{n=1}^{\infty} A_n \times \{y_n\},$$

where each A_n belongs to Σ (The proof of this is a simple extension of arguments given in Chapter 5). We then define $\nu = \nu(c, \mu)$ by

$$\nu\left(\bigcup_{n=1}^{\infty} A_n \times \{y_n\}\right) = \sum_{n=1}^{\infty} \mu\left(c^{-1}(y_n) \cap A_n\right).$$

Then ν is a probability measure, and

Proposition 8.1 *For any $A \in \Sigma$, the sets*

$$A_c = \{(x, c(x)) : x \in A\}$$

and

$$\hat{A}_c = (A \times Y) \setminus A_c = \{(x, y) : x \in A,\, y \neq c(x)\}$$

belong to the σ-algebra $\Sigma \times 2^Y$, and if $\nu = \nu(c, \mu)$ is the measure defined above,

$$\nu(A_c) = \mu(A), \quad \nu(\hat{A}_c) = 0.$$

Proof We have

$$\begin{aligned}\{(x, c(x)) : x \in A\} &= \bigcup_{n=1}^{\infty} \{(x, y_n) : c(x) = y_n \text{ and } x \in A\} \\ &= \bigcup_{n=1}^{\infty} \left(c^{-1}(y_n) \cap A\right) \times \{y_n\},\end{aligned}$$

which is a measurable set. Further,

$$\begin{aligned}\nu(A_c) &= \sum_{n=1}^{\infty} \mu\left(c^{-1}(y_n) \cap \left(A \cap c^{-1}(y_n)\right)\right) \\ &= \sum_{n=1}^{\infty} \mu\left(A \cap c^{-1}(y_n)\right) \\ &= \mu\left(\bigcup_{n=1}^{\infty} A \cap c^{-1}(y_n)\right) \\ &= \mu(A).\end{aligned}$$

Similarly,

$$\hat{A}_c = \{(x, y) : x \in A,\, y \neq c(x)\} = \bigcup_{n=1}^{\infty} \left(A \setminus c^{-1}(y_n)\right) \times \{y_n\},$$

which is a measurable set. Also,

$$\nu(\hat{A}_c) = \sum_{n=1}^{\infty} \mu\left(c^{-1}(y_n) \cap \left(A \setminus c^{-1}(y_n)\right)\right) = 0.$$

□

As in Chapter 5, we shall call $\nu(c, \mu)$ the *deterministic concept representing c and* μ.

When Y is uncountable, we cannot necessarily find $\nu = \nu(c, \mu)$ which represents μ and c in the above sense. However, we can still go some way towards it, as we now show.

Suppose that the uncountable set Y has the σ-algebra $\mathcal{B}$ defined on it. (Usually, we think of Y as some subset of Euclidean space, and $\mathcal{B}$ as the induced Borel sigma-algebra.) Let c be a function from X to Y. Then the function

$$F : X \to X \times Y$$

defined by

$$F(x) = (x, c(x))$$

is $(\Sigma, \Sigma \times \mathcal{B})$-measurable. To show this, we need only verify that

$$A \in \Sigma,\ B \in \mathcal{B} \Longrightarrow F^{-1}(A \times B) \in \Sigma.$$

Now,

$$F^{-1}(A \times B) = \{x \in X : (x, c(x)) \in A \times B\} = A \cap c^{-1}(B) \in \Sigma,$$

and so F is measurable. We now define the measure $\nu = \nu(c, \mu)$ on $\Sigma \times \mathcal{B}$ by

$$\nu(E) = \mu\left(F^{-1}(E)\right), \quad E \in \Sigma \times \mathcal{B}.$$

Then ν has the property that for any $A \in \Sigma$ and any $B \in \mathcal{B}$,

$$\nu(A \times B) = \mu\left(A \cap c^{-1}(B)\right),$$

and in this sense $\nu(c, \mu)$ represents the target concept c together with the underlying distribution μ on X.

Approximating stochastic concepts of countable range

When Y is countable, for any $h \in H$ and any probability measure ν on the σ-algebra $\Sigma \times 2^Y$ (that is, any stochastic concept with range Y defined on X), the set

$$\{(x, y) : y \neq h(x)\}$$

is measurable. For,

$$\begin{aligned}\{(x, y) : y \neq h(x)\} &= (X \times Y) \setminus \{(x, y) : y = h(x)\} \\ &= (X \times Y) \setminus \bigcup_{n=1}^{\infty} h^{-1}(y_n) \times \{y_n\},\end{aligned}$$

which is measurable by the elementary properties of a σ-algebra. We can, therefore, as the obvious extension to a definition of Chapter 5, define the actual error of a hypothesis $h \in H$ to be the measure of this set. A training sample of a stochastic concept and the observed error of a hypothesis on the training sample can be defined as in Chapter 5.

Learnability results

In the (non-stochastic) standard learning framework of Chapter 1, we have a target concept $c : X \to \{0, 1\}$ to be learned. The same notation can be extended to discuss the learnability of a target concept $c : X \to Y$ where Y is an arbitrary set. In the obvious manner, for $h \in H$, μ a probability measure on X, and $\mathbf{x} \in X^m$, we can define $\mathrm{er}_\mu(h)$ and $H[\mathbf{x}]$.

Applying the standard learnability result, Theorem 3.15, to the problem of learning a function with countable range (re-interpreted as the problem of learning the graph of the function), we have the following.

Theorem 8.2 *Let $0 < \epsilon, \delta < 1$ and suppose H is a hypothesis space of finite (generalized) VC dimension $d > 1$ of functions from an input space X to a countable set Y. Let $c \in H$ be any target concept and μ any probability measure on X. Then*

$$\mu^m \{\mathbf{x} \in X^m : \mathrm{haz}_\mu(H[\mathbf{x}]) > \epsilon\} < \delta$$

for

$$m \geq m_0(\epsilon, \delta) = \left\lceil \frac{1}{\epsilon(1-\sqrt{\epsilon})} \left(2d \log \left(\frac{2e}{\epsilon} \right) + \log \left(\frac{d/(d-1)}{\delta} \right) \right) \right\rceil .$$

Proof Let $c \in H$ be any target concept, and

$$\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m.$$

For ease of notation, identify $\mathcal{G}(h)$ and its characteristic function, for any $h \in H$. Then

$$\begin{aligned}
& h \in H[\mathbf{x}, c] \\
\iff & h(x_i) = c(x_i) \, (1 \leq i \leq m) \\
\iff & \mathcal{G}(h)(x_i, c(x_i)) = \mathcal{G}(c)(x_i, c(x_i)) \ (= 1) \, (1 \leq i \leq m) \\
\iff & \mathcal{G}(h) \in \mathcal{G}(H)[c(\mathbf{x}), \mathcal{G}(c)],
\end{aligned}$$

where, as in Chapter 1,

$$c(\mathbf{x}) = ((x_1, c(x_1)), (x_2, c(x_2)), \ldots, (x_m, c(x_m))) .$$

Now, by Proposition 8.1, if $\nu = \nu(c, \mu)$, we have

$$\begin{aligned}
\mathrm{er}_\nu(\mathcal{G}(h), \mathcal{G}(c)) &= \nu \{(x, y) \in X \times Y : c(x) = y \neq h(x) \ \text{ or } \ h(x) = y \neq c(x)\} \\
&= \nu \{(x, c(x)) : h(x) \neq c(x)\} + \nu \{(x, y) : h(x) = y \neq c(x)\} \\
&= \mu \{x : h(x) \neq c(x)\} \\
&= \mathrm{er}_\mu(h, c).
\end{aligned}$$

It follows, since $\mathcal{G}(H)$ has VC dimension d, that with m_0 as in the statement of the theorem, if $m \geq m_0$,

$$\begin{aligned}
& \mu^m \{\mathbf{x} \in X^m : \mathrm{haz}_\mu(H[\mathbf{x}], c) > \epsilon\} \\
&= \mu^m \{\mathbf{x} \in X^m : \mathrm{haz}_\nu \left(\mathcal{G}(H)[c(\mathbf{x}), \mathcal{G}(c)]\right) > \epsilon\} \\
&= \nu^m \{c(\mathbf{x}) \in (X \times Y)^m : \mathrm{haz}_\nu \left(\mathcal{G}(H)[c(\mathbf{x}), \mathcal{G}(c)]\right) > \epsilon\} < \delta.
\end{aligned}$$

The result follows. □

Similarly, we have the following result concerning the approximation of stochastic concepts by hypotheses from a space of finite VC dimension.

Theorem 8.3 *Let $0 < \epsilon < 1$ and $0 < \gamma \leq 1$. Suppose H is a hypothesis space of functions from an input space X to a countable set Y, and let ν be any probability measure on $S = X \times Y$ ~~and $c \in H$ any target concept.~~ Then the probability (with respect to ν^m) that, for $\mathbf{s} \in S^m$, there is some $h \in H$ such that*

$$\mathrm{er}_\nu(h) > \epsilon \quad \text{and} \quad \mathrm{er}_\mathbf{s}(h) \leq (1-\gamma)\mathrm{er}_\nu(h)$$

is at most

$$4\,\Pi_H(2m)\exp\left(-\frac{\gamma^2\epsilon m}{4}\right).$$

Furthermore, if H has finite (generalized) VC dimension d, this quantity is less than δ for

$$m \geq m_0(\epsilon,\delta,\gamma) = \left\lceil \frac{1}{\gamma^2\epsilon(1-\sqrt{\epsilon})}\left(4\log\left(\frac{4}{\delta}\right) + 6d\log\left(\frac{4}{\gamma^{2/3}\epsilon}\right)\right)\right\rceil.$$

Proof The proof of this is similar to the proof of the parallel result, Theorem 5.8, for Boolean stochastic concepts. As there, let the error set E_h of $h \in H$ be the set

$$E_h = \{(x,y) \in X \times Y : h(x) \neq y\}.$$

Observe that $E_h = (X \times Y) \setminus \mathcal{G}(h)$. Let

$$\mathcal{C} = \{E_h : h \in H\}$$

be the collection of error sets. Then the growth function of $\mathcal{C}$ is the same as the growth function of the graph space $\mathcal{G}(H)$. Indeed, suppose that $h, g \in H$, let

$$\mathbf{s} = ((x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)) \in (X \times Y)^m,$$

and let

$$I(\mathbf{s}) = \{(x_1,y_1),(x_2,y_2)\ldots,(x_m,y_m)\}.$$

Then we have, using the fact that E_h is the complement of $\mathcal{G}(h)$,

$$\begin{aligned} & E_h \cap I(\mathbf{s}) = E_g \cap I(\mathbf{s}) \\ \iff & I(\mathbf{s}) \setminus \mathcal{G}(h) = I(\mathbf{s}) \setminus \mathcal{G}(g) \\ \iff & \mathcal{G}(h) \cap I(\mathbf{s}) = \mathcal{G}(g) \cap I(\mathbf{s}). \end{aligned}$$

Therefore, as in the proof of Lemma 5.7, $\Pi_H(\mathbf{s}) = \Pi_{\mathcal{C}}(\mathbf{s})$, and $\Pi_H(m) = \Pi_{\mathcal{C}}(m)$. The proof now proceeds as in the proof of Theorem 5.8, on noting that the VC dimension of the graph space is, by definition, d. □

Chapter 9
An Application to Artificial Neural Networks

9.1 Introduction

In this chapter, we apply the results we have obtained on the learnability of functions to the important problem of bounding the number of training examples which should be presented to an artificial neural network. We discuss previous results for particular families of artificial neural networks and then obtain an upper bound on sufficient sample-size for learning in multiple output feedforward linear threshold nets with real-valued inputs [5, 30]. Our result improves upon a result of Natarajan for multiple output feedforward linear threshold networks with Boolean inputs, by a factor at least equal to the number of nodes in the network. Further, it is more general, applying to the case in which the inputs can be arbitrary real numbers. The bound depends only on the number of nodes and weights in the network and generalizes a result Baum and Haussler obtained for single output feedforward threshold networks with real-valued inputs.

9.2 Artificial Neural Networks

Artificial neural networks [28, 9] have recently received much attention. In particular, many researchers are involved in studying the problem of training a network to compute particular functions and to generalize from the training data presented to it. Before addressing this problem, we have to define mathematically what we mean by an artificial neural network. As the name suggests, these are computational systems in which the method of computation and transfer of information in some way reflects the neuronal structure

of the brain. Here, we discuss the computational or representational power of particular types of network and do not enter the debate on whether artificial neural networks are any valid model of real brains. It is for this reason that we use the word *artificial.*

Definitions

The networks we describe are *feedforward* networks. Roughly speaking, these are networks with no cycles or feedback. A feedforward neural network is an ordered pair

$$\mathcal{N} = (G, \mathcal{F}),$$

where $G = (V, E)$ is a directed acyclic graph and $\mathcal{F}$ is a finite set of *activation functions.* V is the disjoint union of a set I of *input nodes* and a set C of *computation nodes*, and $O \subseteq C$ is the set of *output nodes.* Further, there is a *bias node* $n_0 \in I$. The number of input nodes will be denoted $s + 1$ and the number of output nodes t. The underlying graph G is such that all computation nodes are connected to the bias node and the input nodes have zero in-degree. That is,

$$E \subseteq (C \cup I) \times C$$

and

$$\{n_0\} \times C \subseteq E.$$

The computation nodes are labelled with the integers 1 to $n = |C|$ in such a way that

$$(i, j) \in E \Longrightarrow j > i.$$

This can be accomplished since G is acyclic. We denote by $d(j)$ the in-degree of computation node j.

Associated with computation node j is a set of states $\Omega_j \subseteq \mathbf{R}^{d(j)}$. We let $\Omega^{(k)}$ denote the product

$$\Omega^{(k)} = \Omega_1 \times \ldots \times \Omega_k,$$

and we denote $\Omega^{(n)}$ simply by Ω (this is the set of all states of the network). Any $\omega \in \Omega$ can be decomposed as

$$\omega = \omega_1 \omega_2 \ldots \omega_n,$$

where, for each i between 1 and n, $\omega_i \in \Omega_i$. Given such a decomposition, we denote by ω^k the vector $\omega_1 \omega_2 \ldots \omega_k$.

One thinks of a state of the network as describing the *weights*, or connection strengths, on the edges of the underlying directed graph. In particular, we think of Ω_j as the set of all possible allowed weights on the edges into node j. We use W to denote the number of weights in the network; thus, $W = |E|$.

Each computation node j has associated with it an *activation function*

$$f^j : \Omega_j \times \mathbf{R}^{d(j)} \to \mathbf{R},$$

and $\mathcal{F}$ is the set of n activation functions. Writing $\omega = \omega_j$, the function

$$h^j_\omega : \mathbf{R}^{d(j)} \to \mathbf{R}$$

is given by

$$h^j_\omega(y) = f^j(\omega, y),$$

and we let H^j denote the set of functions h^j_ω where ω runs through Ω_j.

An input $x \in \mathbf{R}^s$ to the network consists of an assignment of a real number to each non-bias input node. Further, each node has an output value, this being defined recursively in terms of the outputs of the previous nodes. The output of a non-bias input node is defined to be the input on that node, and the output of n_0 is always 1. The *input vector* to computation node j depends on the input x and on ω^{j-1}, and we write it as $\mathbf{I}_j(\omega^{j-1}, x) \in \mathbf{R}^{d(j)}$. The first entry of input vector $\mathbf{I}_j(\omega^{j-1}, x)$ is 1, representing the fixed input to the node from the bias node; the other entries represent the (variable) outputs from the nodes adjacent to node j. The output of node j is then computed as

$$f^j\left(\omega_j, \mathbf{I}_j(w^{j-1}, x)\right).$$

The function computed by the network when in state $\omega \in \Omega$ is the function F_ω from $\mathbf{R}^s$ to $\mathbf{R}^t$ whose value is the vector of outputs of the output nodes. The set of all F_ω as ω ranges through Ω is denoted $F(\mathcal{N})$, and we call $F(\mathcal{N})$ the set of functions computable by $\mathcal{N}$.

A network is said to be a *layered network* with h hidden layers if the nodes of the network can be decomposed into $h+2$ sets called *layers* such that layer 0 is the set of input nodes, layer h is the set of output nodes, and if $(i,j) \in E$ (that is, node i is connected to node j) then there is some k such that i belongs to layer k and j to layer $k+1$. Thus the network is feedforward and the only connections are from one layer to the next.

Types of feedforward artificial neural network

We now describe two basic types of feedforward artificial neural network which have been studied theoretically and used in practice.

Perhaps the most general class of neural network consists of networks with real-valued activation functions which are evaluated by adding some function of the inputs to a node j with the weighted sum of the outputs of the nodes connected to j, and passing the result through some suitably well-behaved monotonic real function. Suppose that the node j has in-degree d and denote the inner product of the vectors $\mathrm{y}, \mathrm{z} \in \mathbf{R}^d$ by $\langle \mathrm{y}, \mathrm{z} \rangle$. Suppose that the activation function at node j is such that the output of node j is of the form

$$f^j(\omega_j, \mathbf{I}_j) = \sigma_j\left(\mu_j(\mathbf{I}_j) + \langle \omega_j, \mathbf{I}_j \rangle\right),$$

where

$$\mu_j : \mathbf{R}^d \to \mathbf{R}$$

is a fixed Lipschitz continuous function and

$$\sigma_j : \mathbf{R} \to [0,1]$$

is an arbitrary non-decreasing (or non-increasing) Lipschitz continuous function. The function μ_j is known as the *modifier* for node j. We shall call the

function σ_j the *through-function* of node j. It is often assumed that σ_j is differentiable, but we shall only require that it is Lipschitz continuous for the results we describe. In the literature, σ_j is often called the activation function of node j, but this is not appropriate in our framework. If every activation function of the network is of this form, with a Lipschitz continuous σ_j, we shall say that the network is a ***Lipschitz network.***

Recall that to say a real-valued function σ is Lipschitz continuous on a region D of some Euclidean space means that there is some constant K, a ***Lipschitz bound for*** f ***on*** D, such that for any $x, y \in D$,

$$|f(x) - f(y)| \leq K\,|x - y|\,.$$

We now describe another major class of networks; the *feedforward linear threshold networks.* We say that $\mathcal{N}$ is a feedforward linear threshold network in the case when, for each j between 1 and n, the activation function f^j computes the weighted sum of the outputs of the nodes adjacent to node j and outputs 1 if this is non-negative and 0 otherwise.

This can be described in a manner similar to that used to describe Lipschitz networks. The output of node j is defined to be

$$f^j(\omega_j, \mathbf{I}_j) = \sigma\left(\langle \omega_j, \mathbf{I}_j \rangle\right),$$

where the through-function σ is the linear threshold step function which has value 1 if its argument is non-negative and value 0 otherwise. Clearly, the linear threshold networks are not Lipschitz networks, as the linear threshold function is not Lipschitz continuous in any region containing the origin.

We shall call a neural network consisting of a single linear threshold a *perceptron* (Minsky and Papert [22] give a more general definition of a perceptron). More generally, for this reason, a layered linear threshold network is often described as a *multilayered perceptron.*

Learnability in artificial neural networks

An artificial neural network is trained to compute a function of its inputs by presenting certain training examples together with the required output on these examples. As earlier, regarding this as a sequence of labelled inputs, we call the set of examples a training sample. The state of the network is changed by some means so that the function computed by the network agrees with the target function on all, or on a large fraction of, the sample. Thus, learning in such a system consists of changing the connection weights in response to the presentation of training examples. Many learning algorithms, both on-line algorithms and batch-processing algorithms, have been investigated and implemented in particular families of network. For example, we have the back-propogation algorithm [28] and the linear programming algorithm [31] for layered Lipschitz networks in which the through-functions are differentiable, and the perceptron learning algorithm [28, 22, 9] for perceptrons. We shall not address the problem of finding efficient learning algorithms. This is a difficult issue; indeed, there is complexity-theoretic evidence for the non-existence of successful learning algorithms in many cases [10]. Here, we shall consider the following important question:

Given an artificial neural network and a learning algorithm for that network, how large a training sample should be used so that the function computed by the network after training is a good approximation to the target function?

That is, how large should the sample be in order that the network achieves a valid level of generalization from the training sample?

We can immediately formulate the problem in the probably approximately correct learnability framework:

Suppose that there is a fixed (but possibly unknown) probability distribution on the set of all possible inputs to the network, and that a training sample is drawn according to this distribution. Given a desired level of accuracy ϵ and a confidence parameter δ, how large should the training sample be in order

that, with probability at least $1-\delta$, the network, after training on the sample, computes a function which with probability at least $1-\epsilon$ computes the correct value on a further randomly chosen input? In particular, can one give an upper bound on this sufficient sample-size which is independent of the distribution?

With the previous theory in mind, we are lead naturally to consider the (generalized) VC dimension of $F(\mathcal{N})$, the space of all functions of the inputs that can possibly be implemented on the network. We call the functions in $F(\mathcal{N})$ the set of functions *realisable* or *computable* by the network.

9.3 Previous Results

In this section, we discuss some previously known results. These give sample-size bounds for learnability in layered Lipschitz networks and in feedforward linear threshold networks with a single output and real-valued inputs. We also describe a result of Natarajan, which can be used to provide sample-size bounds for multiple output (not necessarily feedforward) linear threshold nets in which the inputs are constrained to be either 0 or 1.

Suppose that the artificial neural network has, as above, $t \geq 1$ output nodes, so that the output space Y is (some subset of) $\mathbf{R}^t$ if the outputs are real, and $\{0,1\}^t$ if the outputs are binary. As earlier, in studying learnability it is useful to consider distributions on the set $X \times Y$ rather than pairs (c, μ) where c is some measurable function from X to Y and μ is some probability measure on X. We mention that, as described earlier (subject to some measurability conditions) any such pair can be represented as a distribution on the product space. The same terminology as before will be used, so that we will often describe the set of functions computable by a network as a hypothesis space, and so on.

A result of Haussler on Lipschitz networks

We begin by describing a result due to Haussler [18]. This result uses a different measure of the error of a hypothesis and is relevant for the family of layered Lipschitz networks. It does not translate into a result on linear threshold networks, the main topic of this chapter, because the through-functions must be Lipschitz continuous; indeed, the Lipschitz bounds appear in the sample-size bound. However, a significant observation is that if we use our definition of error, the sufficient sample-size of Haussler translates into one that depends linearly on the number of outputs.

The L_1-metric d_1 on $\mathbf{R}^t$ is defined by

$$d_1\left((y_1, y_2, \ldots, y_t), (z_1, z_2, \ldots, z_t)\right) = \frac{1}{t}\sum_{i=1}^{t} |y_i - z_i| .$$

Haussler defines the error $\mathrm{ER}_\nu(h)$ of a hypothesis h to be the expected value, $\mathbf{E}\left(d_1(f(x), y)\right)$ (with respect to the measure ν on $X \times Y$) of $d_1(f(x), y)$. Let d_0 be the discrete metric on Y, which has value 1 unless its arguments are equal, in which case it has value 0. Then the standard definition of error that we have used up to now can be expressed as the expected value, with respect to ν, of $d_0(f(x), y)$. This is the case simply because the expected value of $d_0(f(x), y)$ is precisely the measure of the set of (x, y) for which $d_0(f(x), y)$ takes the value 1; that is, the probability that $f(x) \neq y$. If the outputs can be real numbers and are not restricted to be 0 or 1, it seems more sensible to use Haussler's definition of error. However, our main concern in this chapter is with networks where the outputs are binary.

The following elementary result relates the metrics d_1 and d_0.

Lemma 9.1 *For any* $\mathrm{y} = (y_1, \ldots, y_t)$ *and* $\mathrm{z} = (z_1, \ldots, z_t)$ *in* $\{0,1\}^t$,

$$\frac{1}{t} d_0(\mathrm{y}, \mathrm{z}) \leq d_1(\mathrm{y}, \mathrm{z}) \leq d_0(\mathrm{y}, \mathrm{z}).$$

Let

$$z = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

be a training sample for a neural network with input space X and output space Y. For any $h \in F(\mathcal{N})$, let $\mathrm{ER}_z(h)$ denote the empirical estimate of $\mathrm{ER}_\nu(h)$ on the sample z. That is,

$$\mathrm{ER}_z(h) = \frac{1}{m} \sum_{i=1}^{m} d_1 \left(h(x_i), y_i \right).$$

This is the observed error in the L_1-metric of the hypothesis h on the training sample.

A special case of Haussler's result can be stated as follows.

Theorem 9.2 *Let $\mathcal{N}$ be a layered Lipschitz network with h hidden layers and let $0 < \epsilon, \delta < 1$. Suppose that the weights are bounded in absolute value by β and that there are at most l nodes in a layer. Suppose further that there is a Lipschitz bound of at most s on each through-function, and that each modifier has a Lipschitz bound of at most r, where these quantities satisfy $s(\beta l + r) \geq 1$. Suppose that ν is some probability measure on $\mathbf{R}^s \times \mathbf{R}^t$ and that a training sample z of length m is drawn according to ν. Then there is a sample-size $m_0 = m_0(\epsilon, \delta)$ such that if $m \geq m_0$ then the probability that there is some h in $F(\mathcal{N})$ with*

$$\frac{|\mathrm{ER}_z(h) - \mathrm{ER}_\nu(h)|}{\mathrm{ER}_z(h) + \mathrm{ER}_\nu(h) + \epsilon} > \frac{1}{2}$$

is at most δ. This sufficient sample-size m_0 satisfies

$$m_0 = O\left(\frac{1}{\epsilon}\left(W\left(\log\left(\frac{1}{\epsilon}\right) + h\log\left(s\left(\beta l + r\right)\right)\right) + \log\left(\frac{1}{\delta}\right)\right)\right),$$

where W is the total number of weights. □

This result looks rather unwieldy, but it has some interesting implications. Suppose, in particular, that the relative error of h on the sample z is required to be 0; that is, the final state of the network after training is such that the

function it computes is consistent with the training sample. The result then shows that for a sample-size $m \geq m_0(\epsilon, \delta)$ of the order detailed in the theorem, the probability that h has actual L_1-error $\mathrm{ER}_\nu(h)$ greater than ϵ is at most δ. Now, by Lemma 9.1, if $\mathrm{er}_\nu(h)$, the expected value of $d_0(h(x), y)$ (our standard measure of actual error) is greater than ϵ then the above error, which is defined to be the expected value of $d_1(h(x), y)$, is greater than ϵ/t. This occurs with probability less than δ for $m \geq m_0(\epsilon/t, \delta)$, where the function m_0 is as in the theorem. Therefore, this result gives a sufficient sample-size which, if we use the discrete metric to measure errors on the output, varies more than linearly with the number of output nodes. Specifically, if all other parameters are fixed, the upper bound on sufficient sample-size that it implies for a network with t outputs and W weights is

$$O\left(\frac{t}{\epsilon}\left(W\log\left(\frac{t}{\epsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right).$$

We mention again that this theorem has no relevance to linear threshold networks, because it involves a Lipschitz bound on the through-functions.

Linear threshold networks

Baum and Haussler [6] considered linear threshold networks with a single output. In this case, the generalized VC dimension is simply the standard VC dimension, as the functions computable by the network have range $\{0, 1\}$. They show the following.

Theorem 9.3 *Let $F(\mathcal{N})$ be the space of functions computable by a feed-forward linear threshold neural network $\mathcal{N}$ with n computation nodes and one output node. Then*

$$\mathrm{VCdim}\,(F(\mathcal{N})) \leq 2\,W\log(en),$$

where W is the number of weights in the network. □

From this result we can, as in previous chapters, obtain a sufficient sample-size for learnability to given accuracy with given confidence. This bound on

the VC dimension of the network depends only on the number of weights and the number of nodes in the network, and hence so does the resulting sample-size bound. That this bound is effectively independent of the structure of the underlying graph suggests that it is not tight. In particular, it is not known whether the logarithmic term is necessary. However, Haussler has reported (personal communication) that recent experimental results of Baum seem to suggest that it may be.

Baum and Haussler have also given a lower bound on the VC dimension of some threshold networks. Specifically, they have shown that a layered feedforward threshold network with s (non-bias) inputs, one hidden layer of k nodes and a single output node has VC dimension at least $2\lfloor k/2 \rfloor s$. Notice that this lower bound has no logarithmic term and is approximately equal to the number of weights in the network.

Natarajan [25] has (essentially) obtained a bound on the VC dimension of linear threshold networks with any number of output nodes and $\{0,1\}$-valued inputs. He uses a result of J.W. Hong (personal communication to Natarajan, 1987). Hong's result is that a (not necessarily feedforward) linear threshold network $\mathcal{N}$ with Boolean-valued inputs, n nodes and weights of arbitrary precision (that is, real weights) can be replaced exactly by a linear threshold network with n nodes and $n \log n$-bit integer weights. Therefore, the number of functions computable by the original network is at most the number of possible assignments of $n \log n$-bit integers to each of the W weights of the network, which is

$$\left(2^{n \log n}\right)^W .$$

Thus

$$|F(\mathcal{N})| \leq 2^{Wn \log n}$$

and hence the graph space of the set of functions computable by the network has at most this cardinality. It follows that

$$\begin{aligned}\text{VCdim}\,(F(\mathcal{N})) &\leq \log_2\left(|\mathcal{G}\,(F\,(\mathcal{N}))|\right)\\ &\leq \log_2 2^{Wn\log n}\\ &= Wn\log n.\end{aligned}$$

Thus,

Theorem 9.4 *If $\mathcal{N}$ is a linear threshold neural network with Boolean-valued inputs, n nodes, W weights, and possibly more than one output, then*

$$\text{VCdim}\,(F(\mathcal{N})) \leq Wn\log n.$$

□

Note that the n of this theorem is the total number of nodes, not merely the number of computation nodes.

In the next section, we prove that the upper bound of Baum and Haussler is in fact also an upper bound on the (generalized) VC dimension of a feedforward linear threshold network with real inputs and more than one output node. For feedforward networks, this betters Theorem 9.4 by a factor at least equal to the number of nodes, and it is more general than that theorem, applying to networks with real-valued inputs.

9.4 Multiple Output Threshold Networks

The output function

In this section, we prove the result described above. In order to do this, we first make some further definitions. Recall that we now consider linear threshold networks, and so the output of any computation node will be either 0 or 1.

The *output function* of the network, which describes precisely the output of each computation node, is the function

$$\sigma : \Omega \times X \to \{0,1\}^n.$$

Entry i of $\sigma(\omega, x)$ is 1 if and only if when the network is in state ω and receives input x, node i has output 1. For a sequence $\mathrm{x} = (x_1, \ldots, x_m)$ of inputs, we define $S(\mathcal{N}, \mathrm{x})$ to be the number of distinct vectors of the form

$$(\sigma(\omega, x_1), \ldots, \sigma(\omega, x_m)),$$

where ω runs through all the states in Ω, and we define $S(\mathcal{N}, m)$ to be the maximum over all $\mathrm{x} \in X^m$ of $S(\mathcal{N}, \mathrm{x})$. Clearly if F denotes $F(\mathcal{N})$ then, using the notation of Chapter 8,

$$\Pi_F(m) \leq \Delta_F(m) \leq S(\mathcal{N}, m).$$

Therefore a bound on $S(\mathcal{N}, m)$ is also a bound on $\Pi_F(m)$. In the following, we simplify the notation and denote $S(\mathcal{N}, \mathrm{x})$ and $S(\mathcal{N}, m)$ by $S(\mathrm{x})$ and $S(m)$ (respectively).

VC dimension of multiple output threshold networks

We bound $S(m)$ in the following lemma, obtaining the same bound as was obtained in [6] for the case of one output. This result is not restricted to the class of threshold networks, but applies more generally to feedforward networks in which each activation function has range $\{0,1\}$. For such networks, we denote by $\Pi_j(m)$ the growth function of the space H^j of Boolean-valued functions computed by the computation node j. With this, we have

Lemma 9.5 *With the above notation, for any positive integer m,*

$$S(m) \leq \Pi_1(m)\Pi_2(m)\ldots\Pi_n(m).$$

Proof For any i between 1 and n, let $\mathcal{N}_i$ be the subnetwork induced by the input nodes and nodes 1 to i, which is itself a feedforward linear threshold network. Observing that the set of states of $\mathcal{N}_i$ is $\Omega^{(i)}$, let

$$\sigma_i : \Omega^{(i)} \times X \to \{0,1\}^i$$

be the output function of $\mathcal{N}_i$. Further, for each i between 1 and m, let $S_i(m) = S(\mathcal{N}_i, m)$ be defined for the network $\mathcal{N}_i$ in the same way as $S(m)$ is defined for $\mathcal{N}$. We claim that for any i between 1 and n,

$$S_i(m) \leq \Pi_1(m)\Pi_2(m)\ldots\Pi_i(m),$$

from which the result will follow. We prove the claim by induction.

The base case is easily seen to be true; $S_1(m) = \Pi_1(m)$, since the output function in this case is exactly the output of node 1.

Assume that the claim holds for $i = k-1$ $(k \geq 2)$ and consider now the case $i = k$. Observe that, writing $\omega \in \Omega^{(k)}$ as

$$\omega = \omega^{k-1}\omega_k,$$

where $\omega^{k-1} \in \Omega^{(k-1)}$ and $\omega_k \in \Omega_k$, we have

$$\sigma_k(\omega^k, x) = \sigma_k(\omega^{k-1}\omega_k, x) = (\sigma_{k-1}(\omega^{k-1}, x), f^k(\omega_k, \mathbf{I}_k(\omega^{k-1}, x)).$$

Let $\mathrm{x} = (x_1, \ldots, x_m) \in X^m$. The number of vectors of the form

$$\left(\sigma_{k-1}(\omega^{k-1}, x_1), \ldots, \sigma_{k-1}(\omega^{k-1}, x_m)\right)$$

as ω^{k-1} ranges through $\Omega^{(k-1)}$ is at most $S_{k-1}(m)$ and, for a fixed ω^{k-1} in $\Omega^{(k-1)}$ the number of vectors of the form

$$\left(f^k\left(\omega_k, \mathbf{I}_k(\omega^{k-1}, x_1)\right), \ldots, f^k\left(\omega_k, \mathbf{I}_k(w^{k-1}, x_m)\right)\right)$$

as ω_k ranges through Ω_k is at most $\Pi_k(m)$. Thus, for any $\mathrm{x} = (x_1, \ldots, x_m)$ in X^m, the number of vectors of the form

$$\left(\sigma_k(\omega^{k-1}\omega_k, x_1), \ldots, \sigma_k(\omega^{k-1}\omega_k, x_m)\right)$$

as $\omega = \omega^{k-1}\omega_k$ ranges through $\Omega^{(k)}$ is at most $\Pi_k(m)S_{k-1}(m)$. Hence

$$\begin{aligned} S_k(m) &\leq \Pi_k(m)S_{k-1}(m) \\ &\leq \Pi_k(m)\Pi_1(m)\Pi_2(m)\ldots\Pi_{k-1}(m) \\ &= \Pi_1(m)\Pi_2(m)\ldots\Pi_k(m), \end{aligned}$$

and the result follows. □

This implies the following extension of a result from [6], which again applies to a general network in which each activation function is Boolean-valued.

Proposition 9.6 *Let $\mathcal{N}$ be a feedforward artificial neural network with real-valued inputs, possibly more than one output, and n computation nodes which have $\{0,1\}$-valued activation functions. Suppose that the VC dimension of the set of functions computable by computation node j is r_j, and let*

$$R = \sum_{j=1}^{n} r_j.$$

Then, for $m \geq R$, if F denotes the set of functions $F(\mathcal{N})$ computable by the network, we have, for $m > R$,

$$\Pi_F(m) \leq \left(\frac{nem}{R}\right)^R.$$

Proof We use the above lemma. Certainly, $R \geq r_j$ for j between 1 and n, and so, for each such j and for $m > R$,

$$\Pi_j(m) \leq \left(\frac{em}{r_j}\right)^{r_j}.$$

It follows from the above result that

$$\begin{aligned}\Pi_F(m) &\leq \Pi_1(m)\Pi_2(m)\ldots\Pi_n(m) \\ &\leq \prod_{j=1}^{n}\left(\frac{em}{r_j}\right)^{r_j}.\end{aligned}$$

Now, if $\alpha_i > 0$ for $1 \leq i \leq n$ and $\sum_{i=1}^n \alpha_i = 1$ then

$$\sum_{i=1}^{n} -\alpha_i \log \alpha_i \leq \log n.$$

Setting $\alpha_i = r_i/R$, we obtain

$$\begin{aligned}&\sum_{i=1}^{n}\frac{r_i}{R}\log\left(\frac{R}{r_i}\right) \leq \log n \\ &\iff \sum_{i=1}^{n} r_i \log\left(\frac{1}{r_i}\right) \leq R\log n - \left(\sum_{i=1}^{n} r_i\right)\log R = R\log n - R\log R \\ &\iff \prod_{i=1}^{n}\left(\frac{1}{r_i}\right)^{r_i} \leq \left(\frac{n}{R}\right)^R,\end{aligned}$$

from which the result follows. □

Recall that in a feedforward linear threshold network, the activation function $f_j \in \mathcal{F}$ computes the inner product of its arguments and outputs 1 if this is non-negative and 0 otherwise. That is, each computation node j computes (some restriction of the characteristic function of) a positive half-space of $\mathbf{R}^{d(j)}$. Thus, by the discussion after Theorem 2.8, the VC dimension of H^j is at most $d(j)$.

Corollary 9.7 *Let $F = F(\mathcal{N})$ be the space of functions computable by a feedforward linear threshold neural network $\mathcal{N}$ with n computation nodes and possibly more than one output node. Then*

$$\mathrm{VCdim(F)} \leq 2W\log(en),$$

where W is the total number of weights in the network.

Proof We use the above Proposition. As discussed above, the VC dimension of H^j is at most $d(j)$. Then

$$R \leq \sum_{i=1}^{n} d(j) = W,$$

the total number of weights in the network. By Proposition 9.6, with $F = F(\mathcal{N})$, for $m \geq W$ we have

$$\Pi_F(m) \leq \left(\frac{nem}{W}\right)^W.$$

Now,

$$\left(\frac{2enW\log(en)}{W}\right)^W < 2^{2W\log(en)}$$
$$\iff 2en\log(en) < (en)^2$$
$$\iff 2\log(en) < en,$$

which is true for any $n \geq 1$. Therefore, $\Pi_F(m) < 2^m$ when $m = 2W\log(en)$, and the VC dimension of F is at most $2W\log(en)$, as required. □

In particular, the VC dimension of the network can be bounded independently of the number of output nodes.

Sample-size bounds

This result has the following immediate implication for generalization in such networks, which applies to the case when there is some function being learned, the training sample is drawn according to some fixed distribution on the input space, and the learning process produces a hypothesis (or state) consistent with the target function on the sample.

Corollary 9.8 *Suppose we are given an accuracy parameter* $0 < \epsilon < 1$ *and a confidence parameter* $0 < \delta < 1$. *Let* $\mathcal{N}$ *be a feedforward linear threshold artificial neural network with* W *variable weights,* n *computation nodes and possibly more than one output node. Suppose that* $\mathcal{N}$ *is being trained to compute some function of its inputs. (We assume that the network is capable of computing this function). Then there is a sample-size* $m_0 = m_0(\epsilon, \delta)$ *such that if* $\mathcal{N}$ *is trained to compute the correct output on a training sample of* $m \geq m_0$ *inputs, chosen according to some distribution on the set of all inputs, then the following holds with probability at least* $1 - \delta$*: With probability at least* $1 - \epsilon$*, for a randomly chosen input, the network computes the correct output. A suitable value of* m_0 *is*

$$m_0 = \left\lceil \frac{1}{\epsilon(1 - \sqrt{\epsilon})} \left(\log\left(\frac{2}{\delta}\right) + 4W \log(en) \log\left(\frac{6}{\epsilon}\right) \right) \right\rceil .$$

□

More generally, when we allow a certain degree of error during the training, or when the learning process need not produce a hypothesis consistent with the training sample but only highly consistent, as in Section 5.4, we have the following.

Corollary 9.9 *With* $\epsilon, \delta, \mathcal{N}, W$ *and* n *as above, suppose that* $0 < \gamma \leq 1$. *Suppose that* $\mathcal{N}$ *is being trained to compute some (computable) target function of its inputs. Then there is* $m_0 = m_0(\epsilon, \delta, \gamma)$ *such that if* $\mathcal{N}$ *is trained to compute the correct output on at least a proportion* $1 - (1 - \gamma)\epsilon$ *of* $m \geq m_0$

inputs, chosen according to some distribution on the set of all inputs, then the following holds with probability at least $1 - \delta$*: With probability at least* $1 - \epsilon$*, for a randomly chosen input, the network computes the correct output. A suitable value of* m_0 *is*

$$m_0(\epsilon, \delta, \gamma) = \left\lceil \frac{1}{\gamma^2 \epsilon (1 - \sqrt{\epsilon})} \left(4 \log \left(\frac{4}{\delta} \right) + 12W \log(en) \log \left(\frac{4}{\gamma^{2/3} \epsilon} \right) \right) \right\rceil .$$

□

References

[1] D. Angluin, Queries and concept learning, *Machine Learning,* 2: 319-342, (1987)

[2] D. Angluin and P. Laird, Learning from noisy examples, *Machine Learning,* 2: 343-370, (1987)

[3] D. Angluin and L.G. Valiant, Fast probabilistic algorithms for Hamiltonian circuits and matchings, *J. Comput. Syst. Sci.,* 19: 155-193, (1979)

[4] M. Anthony, N.L. Biggs, and J. Shawe-Taylor, The learnability of formal concepts, in *COLT 90, Proceedings of the Third Annual Workshop on Computational Learning Theory*, 246-257, Morgan Kaufmann: San Mateo, CA (1990)

[5] M. Anthony and J. Shawe-Taylor, *A result of Vapnik with applications*, Technical Report CSD-TR-628, Department of Computer Science, Royal Holloway and Bedford New College (University of London), July 1990.

[6] E.B. Baum and D. Haussler, What size net gives valid generalization, *Neural Computation,* 1(1): 151-160 (1989)

[7] S. Ben-David, G.M. Benedek, and Y. Mansour, A parameterization scheme for classifying models of learnability, in *COLT 89, Proceedings of the Second Annual Workshop on Computational Learning Theory*, 285-302, Morgan Kaufmann: San Mateo, CA (1989)

[8] G.M. Benedek and A. Itai, Learnability by fixed distributions, in *COLT 88, Proceedings of the First Annual Workshop on Computational Learning Theory*, 80-90, Morgan Kaufmann: San Mateo, CA (1988)

[9] N.L. Biggs, Combinatorics and connectionism, *Discrete Math.,* (to appear)

[10] A. Blum and R.L. Rivest, Training a 3-node neural network is NP-complete, in *COLT 88, Proceedings of the First Annual Workshop on Computational Learning Theory*, 9-18, Morgan Kaufmann: San Mateo, CA (1988)

[11] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, Learnability and the Vapnik-Chervonenkis Dimension, *J. ACM,* 36(4): 929-965, (~~1990~~) 1989

[12] H. Chernoff, A measure of asymptotic efficiency for tests based on the sum of observations, *Annals of Math. Stat.,* 23: 493-507, (1952)

[13] A. Ehrenfeucht, D. Haussler, M. Kearns and L.G. Valiant, A general lower bound on the number of examples needed for learning, *Information and Computation,* 82, 247-261, (1989)

[14] G. Grimmett and D. Welsh, *Probability, an Introduction,* Oxford University Press: Oxford, (1986)

[15] Grunbaum, *Convex Polytopes,* John Wiley: London (1967)

[16] D. Haussler, Quantifying inductive bias: AI learning algorithms and Valiant's learning framework, *Artificial Intelligence,* 36(2): 177-222 (1988)

[17] D. Haussler, Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence results, *unpublished preliminary extended abstract for COLT 89,* 1989

[18] D. Haussler, *Generalizing the PAC model for neural net and other learning applications*, Technical Report UCSC-CRL-89-30, University of California Computer Research Laboratory, Santa Cruz, CA, 1988

[19] D. Haussler and E. Welzl, ϵ-nets and simplex range queries, *Discrete Comp. Geometry,* 2: 127-151 (1987)

[20] P. Laird, *Learning from good data and bad*, Technical Report TR-551, Department of Computer Science, Yale University, 1987

[21] C. McDiarmid, On the method of bounded differences, in *Surveys in Combinatorics, 1989* (ed. J. Siemons), 148-184, Cambridge University Press: Cambridge (1989)

[22] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry,* MIT Press: Cambridge, MA (1988)

[23] J.W. Moon, Four combinatorial problems, in *Combinatorial Mathematics and its Applications* (ed. D.J.A. Welsh), 185-190, Academic Press (1971)

[24] S. Muroga, *Threshold Logic and its Applications,* Wiley: New York (1971)

[25] B.K. Natarajan, On learning sets and functions, *Machine Learning,* 4: 67-97, (1989)

[26] L. Pitt and L.G. Valiant, Computational limits on learning from examples, *J. ACM,* 35(4): 965-984, (1988)

[27] D. Pollard, *Convergence of Stochastic Processes,* Springer: New York (1984)

[28] D. Rumelhart and J.L. McLelland , *Parallel Distributed Processing, Volumes 1 and 2,* MIT Press: Cambridge, MA (1986)

[29] N. Sauer, On the density of families of sets, *J. Combinatorial Theory (A),* 13: 145-147, (1972)

[30] J. Shawe-Taylor and M. Anthony, Sample sizes for multiple output threshold networks, *Network,* (to appear)

[31] J. Shawe-Taylor and D. Cohen, The linear programming algorithm for neural networks, *Neural Networks,* 3: 575-582 (1990)

[32] R. Sloan, Types of noise for concept learning, in *COLT 88, Proceedings of the First Annual Workshop on Computational Learning Theory*, 91-96, Morgan Kaufmann: San Mateo, CA (1988)

[33] L.G. Valiant, A theory of the learnable, *Comm. ACM,* 27(11): 1134-1142 (1984)

[34] L.G. Valiant, Deductive learning, *Phil. Trans. Royal Soc.,* A312: 441-446 (1984)

[35] V. Vapnik, *Estimation of dependences based on Empirical Data,* Springer Verlag: New York (1982)

[36] V.N. Vapnik and A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theor. Probability and Appl.,* 16(2): 264-280, (1971)

[37] R. Wille, Restructuring lattice theory: an approach based on hierarchies of concepts, in *Ordered Sets, NATO Advanced Study Institute Series 83, 445-470, (1982)*